Proceedings of the 18th Seminar on Ontology Research in Brazil and 9th Workshop on Theses and Dissertations in Ontologies

Edited by

Fernanda Farinelli
Joel Luís Carbonera
Amanda Damasceno de Souza
Fabrício Henrique Rodrigues

2025

IAOA
The
International
Association for
Ontology and
its Applications

ONTOBRAS

SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Workshop on Theses and Dissertations in Ontologies (WTDO 2025)

Hosted by: Instituto Tecnológico de Aeronáutica (ITA) e Faculdade de Ciências Médicas Humanitas, São José dos Campos, SP

Brazil, September 29th to October 2nd 2025

CEUR Workshop Proceedings
ceur-ws.org
ISSN 1613-0073

I A O A
The International Association for Ontology and its Applications

Organization:

Support:

The International Association for Ontology and its Applications

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Workshop on Theses and Dissertations in Ontologies (WTDO 2025)

Online publication:

CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org
ISSN publication series: 1613-0073
DOI: 10.5281/zenodo.17508451

Preferred citation:

FARINELLI, F; CARBONERA, J.L; SOUZA, A.D.; RODRIGUES, F.H. (eds.). ONTOBRAS 2025 – 18th Seminar on Ontology Research in Brazil. *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Workshop on Theses and Dissertations in Ontologies (WTDO 2025),* September 29th - October 2nd, 2025. São José dos Campos, SP, Brazil:CEUR-WS.org, 2025.

It is recommended to include further the CEUR-WS.org volume number and the URL to that volume.

Program Chairs' addresses:

Fernanda Farinelli
e-mail: fernanda.farinelli@unb.br
University of Brasília (UnB)
Faculty of Information Sciences
Postal Code 70297-400, Brasilia/DF, Brazil

Joel Luís Carbonera
e-mail: joel.carbonera@inf.ufrgs.br
Federal University of Rio Grande do Sul (UFRGS)
Institute of Informatics
Postal Code 91501-970, Porto Alegre/RS, Brazil

WTDO Chairs' addresses:

Amanda Damasceno de Souza
e-mail: amanda.dsouza@fumec.br
Minas Gerais Foundation for Education and Culture (FUMEC)
Graduate Program in Information and Communication Technology and Knowledge Management
Postal Code 30310-150, Belo Horizonte/MG, Brazil

Fabrício Henrique Rodrigues
e-mail: fr33@buffalo.edu
University at Buffalo
Department of Philosophy
Postal Code 14260-4150, Buffalo, NY, USA
CEP 91501-970, Porto Alegre/RS, Brazil

Preface

Ontology is an interdisciplinary domain that investigates the principles, models, and theories used to construct shared conceptual representations of specific areas of knowledge. Over the past years, interest in ontology has expanded significantly, particularly due to its relevance in addressing modeling, integration, and classification challenges across fields such as Computer Science, Information Science, Artificial Intelligence, Philosophy, Linguistics, and Knowledge Management, among others.

The *Seminar on Ontology Research in Brazil (ONTOBRAS)* provides an academic space and scientific environment for researchers and practitioners from these and related fields to exchange ideas and discuss theories, methodologies, languages, tools, and practical experiences concerning ontology design, development, and application.

In its 18th edition, **ONTOBRAS 2025** continues a tradition established in 2008 in Brazil and recognized by the Brazilian Ontology Community as the main scientific forum for presenting and discussing ontologies and their application cases in the country.

This edition of ONTOBRAS focuses on the theme **"The Influence of Ontologies in Generative AI."** It is a timely and relevant topic for both academia and the productive sector. Recent studies have shown how ontologies and knowledge graphs contribute to improving the quality and consistency of responses generated by artificial intelligence systems. The theme also aims to strengthen the connection between academic research and industry, fostering collaboration, innovation, and the development of new applications and products.

Participants were invited to submit theoretical, methodological, and applied research papers that are directly or indirectly related to ontology studies, written in Portuguese, English, or Spanish. Submissions were organized into two tracks: the *Main Track* and the *Workshop on Theses and Dissertations in Ontologies (WTDO).*

All submissions underwent a double-blind review process conducted by our Program Committee, composed of national and international experts in the field. In the *WTDO*, 6 submissions (4 at the doctoral level and 2 at the master's level) were selected for oral presentation during the event. In total, the *Main Track* received 48 submissions. After the review process, 12 papers were accepted as full papers and 12 as short papers, both for oral presentation and publication in these proceedings. In addition, 11 papers were accepted for poster presentations. During the conference, 11 full papers, 10 short papers, and 5 posters were effectively presented, demonstrating the diversity and quality of the research shared within the ONTOBRAS 2025 community.

The scientific program also featured five keynote talks delivered by renowned national and international experts: Barry Smith (University at Buffalo), Giancarlo Guizzardi (University of Twente), John Beverley (University at Buffalo), Jérémy Ravenel (naas.ai), and Renata Wassermann (University of São Paulo). Additionally, Aaron Damiano (U.S. Customs and Border Protection) presented the ontology demo "Fandaws (Fact and Answer Web Service)". The event also included two tutorials focusing on the Basic Formal Ontology (BFO) and the Unified Foundational Ontology (UFO), offered by John Beverley and Giancarlo Guizzardi, respectively. These sessions enriched the event by promoting reflection, methodological exchange, and dialogue between academia and practice in the field of ontology.

We hope that the contributions presented in these proceedings will inspire new research, foster collaboration, and continue to strengthen the ontology community in Brazil and beyond.

We thank the organizing committee for their commitment to the success of the event, the authors for their submissions and the program committee for their hard work.

December 2025

Fernanda Farinelli
Joel Luís Carbonera
Amanda Damasceno de Souza
Fabrício Henrique Rodrigues

ONTOBRAS 2025

General Chair

José M. Parente de Oliveira - Aeronautics Institute of Technology, Brazil

ONTOBRAS Program Committee Chairs

Fernanda Farinelli - University of Brasília, Brazil
Joel Luís Carbonera - Federal University of Rio Grande do Sul, Brazil

WTDO Program Committee Chairs

Amanda Damasceno De Souza - FUMEC University, Brazil
Fabrício Henrique Rodrigues - Federal University of Rio Grande do Sul, Brazil

Communication Chairs

Evellin Cristine Souza Cardoso - Federal University of Goiás, Brazil
Rafael Logan de Souza Nobre - Federal Technological University of Paraná, Brazil

Web Chairs

Haroldo Rojas de Souza Silva - Federal University of Rio Grande do Sul, Brazil
Rafael Humann Petry - Federal University of Rio Grande do Sul, Brazil

Steering Committee

Amanda Damasceno De Souza - FUMEC University, Brazil
Claudenir Morais Fonseca - University of Twente, The Netherlands
Daniela Lucas - Federal University of Espírito Santo, Brazil
Daniela Schmidt - University of Évora, Portugal
Eduardo Ribeiro Felipe - Federal University of Itajubá, Brazil
Fabrício Henrique Rodrigues - Federal University of Rio Grande do Sul, Brazil
Fernanda Farinelli - University of Brasília, Brazil
Giancarlo Guizzardi - University of Twente, The Netherlands)
Jeanne Louize Emygdio - PUC Minas Gerais, Brazil)
João Luiz Rebelo Moreira - University of Twente, The Netherlands
Joel Luís Carbonera - Federal University of Rio Grande do Sul, Brazil
José M. Parente de Oliveira - Aeronautics Institute of Technology, Brazil
Luan Garcia - Federal University of Rio Grande do Sul, Brazil
Mara Abel - Federal University of Rio Grande do Sul, Brazil
Monalessa Barcellos - Federal University of Espírito Santo, Brazil
Rita Berardi - Federal University of Technology – Paraná, Brazil
Tiago Prince Sales - University of Twente, The Netherlands
Veruska Zamborlini - Federal University of Espírito Santo, Brazil
Vítor E. Silva Souza - Federal University of Espírito Santo, Brazil

Program Committee

Alcides Lopes
Alexandre Rademaker - *Getulio Vargas Foundation, Brazil*
Amanda Damaceno de Souza - FUMEC University, *Brazil*
Ana Carolina S. Arakaki - *University of Brasília, Brazil*
Camila Z. Aguiar - *Federal University of Espírito Santo, Brazil*
Carlos Marcondes - *Fluminense Federal University, Brazil*
Cassia Trojahn - *University of Toulouse, France*
Claudenir M. Fonseca- *University of Twente, The Netherlands*
Cláudio Gottschalg-Duque - *University of Brasília, Brazil*
Clever Farias - *University of São Paulo, Brazil*
Cristine Griffo - *European Academy of Bolzano, Brazil*
Daniela Lucas Da Silva - *Federal University of Espírito Santo, Brazil*
Eduardo Felipe - *Federal University of Itajubá, Brazil*
Evellin Cardoso - *Federal University of Goiás, Brazil*
Fabricio Henrique Rodrigues - *Federal University of Rio Grande do Sul, Brazil*
Felipe Augusto Arakaki - *University of Brasília, Brazil*
Fernanda Baião - *Pontifical Catholic University of Rio de Janeiro, Brazil*
Fernanda Farinelli - *University of Brasília, Brazil*
Fernando Cruz - *University of Brasília, Brazil*
Fernando Ostuni Gauthier - *Federal University of Santa Catarina, Brazil*
Ferrucio de Franco Rosa - *University Center of Campo Limpo Paulista, Brazil*
Flavio S. Correa Da Silva - *University of São Paulo, Brazil*
Fred Freitas - *Federal University of Pernanbuco, Brazil*
Gabriela Henning - *Technological Institute of Santo Domingo, Dominican Republic*
Giancarlo Guizzardi - *University of Twente, The Netherlands*
Ítalo Oliveira - *University of Twente, The Netherlands*
J. Neil Otte - *Johns Hopkins University, United States*
Jeanne Louize Emygdio - *Pontifical Catholic University of Minas Gerais, Brazil*
João Lima - Federal Senate, Brazil
João Paulo Almeida - *Federal University of Espírito Santo, Brazil*
Joel Carbonera - *Federal University of Rio Grande do Sul, Brazil*
John Beverley - *University at Buffalo, United States*
José M. Parente de Oliveira - *Aeronautics Institute of Technology, Brazil*
Júlio Cesar Dos Reis - *University of Campinas, Brazil*
Júlio Cesar Nardi - *Federal Intitute of Espírito Santo, Brazil*
Lais Salvador - *Federal University of Bahia, Brazil*
Luan Garcia - *Federal University of Rio Grande do Sul, Brazil*
Luciano Heitor Gallegos Marin - *Federal Technological University of Paraná, Brazil*
Luís Ferreira Pires- *University of Twente, The Netherlands*
Mara Abel - *Federal University of Rio Grande do Sul, Brazil*
Marcello Peixoto Bax - *Federal University of Minas Gerais Brazil*

Maria das Graças da Silva Teixeira - *Federal University of Espírito Santo, Brazil*
Mathias Brochhausen - *University of Arkansas for Medical Sciences, United States*
Monalessa Barcellos - *Federal University of Espírito Santo, Brazil*
Pedro Paulo F. Barcelos - *Health Research Infrastructure, The Netherlands*
Rafael Peñaloza - *University of Milano-Bicocca, Italy*
Regina Braga - *Federal University of Juiz de Fora, Brazil*
Renata Guizzardi- *University of Twente, The Netherlands*
Sandro Rama Fiorini - *IBM Research, Brazil*
Silvio Gonnet - *Universidad Tecnológica Nacional, Argentina*
Thiago Henrique Bragato Barros - *Federal University of Rio Grande do Sul, Brazil*
Tiago Prince Sales- *University of Twente, The Netherlands*
Veruska Zamborlini - *Federal University of Espírito Santo, Brazil*
Vitor E. Silva Souza - *Federal University of Espírito Santo, Brazil*
Vivian S. Silva - *Federal University of Rio de Janeiro, Brazil*

Contents

Part I: Full Papers

Part II: Short Papers

Closing Remarks

The *Workshop on Theses and Dissertations in Ontologies (WTDO)* at ONTOBRAS 2025 once again offered a valuable academic space for discussion and mentoring among graduate students and senior researchers. The works presented reflected the diversity of research being developed in Brazil and abroad, combining theoretical reflection, methodological innovation, and applied ontology development.

This edition included six presentations, divided between doctoral *theses* and master's *dissertations*, as follows:

Theses

- *An Ontology-Based Approach to Streamline the Reconstruction of Genome-Scale Metabolic Models*, by Nahim A. Souza and Renata Wassermann.
- *Ontologia de Alto Nível para Segurança Cibernética em Sistemas de Informação*, by Roberto Monteiro Dias and Sean Siqueira.
- *Classification of Gender Stereotypes in a Legal Context: A Systematic Review and Current Opportunities*, by Tainá Turella Caetano dos Santos and Renata Wassermann.
- *Knowledge Management and Organization in Decentralized Finance*, by Fábio Cossenzo and Marcello Bax.

Dissertations

- *Domain Ontology for Mapping Competency Development in Higher Education Engineering Programs*, by Eduardo Perotti, Eduardo Felipe, Giovani Vitor, Fernanda Farinelli, and Rodrigo Braga.
- *OntoMI: An Ontology Grounded in the Theory of Multiple Intelligences for Semantic Classification of Educational Resources*, by Jefferson Rodrigo Speck, Sidgley Camargo de Andrade, and Clodis Boscarioli.

In addition to the WTDO, ONTOBRAS 2025 also featured a **Poster Session**, which provided an open space for sharing innovative ideas, ongoing work, and exploratory studies that connect ontology to emerging technologies and applied contexts. The following posters were presented:

- *Plugin de Alinhamento Interativo de Ontologias de Cibersegurança com Abordagens de Aprendizado de Máquina e K-Means*, by Roberto Dias.
- *ONTOGEN: Ontology-Guided Knowledge Graph Construction for Generative AI*, by Felipe Nunes.
- *The Annotation Jungle: An Analysis of Annotation Property Usage in the Ontology Community*, by Rafael Petry, Haroldo Silva, and Mara Abel.

- *Uma abordagem XAI baseada em Ontologias para Seleção Multicritério de Fornecedores em Cadeias de Suprimento*, by Mateus da Rocha Peixoto, Fernanda Baião, and Renata Guizzardi.
- *Por que a Embrapa precisa de ontologias? Caminhos para a organização e integração do conhecimento agropecuário brasileiro*, by Celina Takemura, Milena Telles, Leandro Oliveira, Bibiana Almeida, Débora Drucker, Jaudete Daltio, Rochelle Alvorcem, and Maria de Cléofas Alencar.
- *Adequação aos princípios FAIR e ontologias biomédicas em instituto de pesquisa neurocientífica: em que medida isso é eficiente na gestão de dados de saúde no Brasil?*, by Arthur Matta and Jonas Reis.

The diversity of topics and institutions represented in the WTDO and Poster Sessions demonstrates the breadth of ontology research and its intersections with artificial intelligence, data governance, education, and knowledge organization. These contributions reaffirm ONTOBRAS's role as a collaborative and interdisciplinary forum that connects theory, technology, and practice, strengthening the ontology research community in Brazil and beyond.

Ontologia para representação do conhecimento sobre consentimento do titular de dados pessoais em mídias sociais digitais

Camila Gourgues Pereira[1,*], Luciano Heitor Gallegos Marin[1] and Cristina Godoy Bernardo de Oliveira[2]

[1] *Federal University of Paraná (UFPR), Prefeito Lothário Meissner Av., 632, 80210-170, Curitiba, PR, Brazil*

[2] *University of São Paulo (USP), Bandeirantes Av., 3900, 14040-906, Ribeirão Preto, SP, Brazil*

Resumo

As mídias sociais digitais transformaram as formas de interação e comunicação na sociedade contemporânea, ao mesmo tempo em que intensificaram a coleta e o uso de dados pessoais. Nesse contexto, uma das bases legais para o tratamento de dados pessoais é o consentimento do usuário, conforme previsto na Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018). No entanto, fatores como a assimetria de poder, a complexidade das políticas de privacidade, a presença de *dark patterns* e os vieses cognitivos dificultam que o consentimento seja realmente livre, informado e inequívoco, conforme previsto em Lei. Esta pesquisa em desenvolvimento, de caráter descritivo, utiliza a metodologia *OntoForInfoScience* para o desenvolvimento de uma estrutura semântica ontológica voltada à representação do consentimento do titular de dados pessoais em mídias sociais digitais. A proposta integra fundamentos legais e aspectos sociocomportamentais, com o objetivo de oferecer uma estrutura semântica visando apoiar soluções tecnológicas para a privacidade de dados pessoais, a conformidade normativa e a conscientização dos usuários.

Palavras-chave

Ontologia, Consentimento, Mídias Sociais Digitais, Proteção de Dados

Abstract

Digital social media has reshaped interaction and communication in contemporary society, while simultaneously intensifying the collection and use of personal data. In this context, consent represents one of the legal bases for processing personal data, as established by the Brazilian General Data Protection Law (Law No. 13,709/2018). However, factors such as power asymmetry, the complexity of privacy policies, the presence of dark patterns, and cognitive biases often prevent consent from being truly free, informed, and unambiguous, as required by law. This descriptive, ongoing research employs the OntoForInfoScience methodology to develop an ontological semantic structure aimed at representing data subjects' consent in digital social media. The proposal integrates legal foundations and socio-behavioral aspects, aiming to provide a semantic framework to support technological solutions for personal data privacy, regulatory compliance, and user awareness.

Keywords

Ontology, Consent, Digital Social Media, Data Protection

1. Introdução

A crescente digitalização da vida social é impulsionada pela convergência das tecnologias da informação e da globalização econômica, que promovem transformações na forma como os indivíduos se relacionam e interagem. Consolida-se, assim, uma nova rede global de comunicação e informação, formada por plataformas de mídias sociais e por dispositivos móveis, em que todos os aspectos da vida social estão interconectados [1].

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

✉ camilagourgues@ufpr.br (C. G. Pereira); luciano.gallegos@ufpr.br (L. H. G. Marin); cristinagodoy@usp.br (C. G. B. Oliveira)

iD 0009-0000-8352-2362 (C. G. Pereira); 0000-0002-4331-6588 (L. H. G. Marin); 0000-0003-4590-3249 (C. G. B. Oliveira)

Nesse contexto, as mídias sociais digitais redefiniram as formas de interação social e tornaram-se centrais na sociedade moderna como espaços de conexão, interação e negócios [2]. Os usuários, ao interagirem com essas tecnologias, recebem e enviam uma grande quantidade de dados, os quais fornecem detalhes de sua vida privada e de suas preferências pessoais. Por conseguinte, essas plataformas formam uma grande rede informacional, na qual circula uma quantidade excessiva de dados e informações a todo momento [1].

Essa dinâmica de fluxo informacional é central na chamada "economia digital", na qual tecnologias de Inteligência Artificial, Internet das Coisas e Big Data dependem do acesso e da análise de dados pessoais para oferecer serviços personalizados. No entanto, tal prática expõe os indivíduos a um monitoramento contínuo que pode infringir a liberdade e gerar risco de vazamento de informações [3]. Como resultado, a coleta e o uso indiscriminado de dados pessoais pelas mídias sociais digitais podem representar riscos à proteção de dados e à privacidade dos usuários [2].

Diante do panorama apresentado, nos últimos anos, a proteção de dados pessoais se consolidou como uma questão global de relevância pública e jurídica. Assim, diversos países promulgaram suas próprias legislações sobre o tema [3]. Como marco importante, pode-se citar o General Data Protection Regulation (GDPR), que entrou em vigor em 2018, o regulamento da União Europeia que afeta qualquer empresa ou organização que trate dados de cidadãos europeus, independentemente de onde esteja localizada [4].

Foi nesse cenário que, no Brasil, a Lei nº 13.709/2018, Lei Geral de Proteção de Dados (LGPD), foi sancionada em 2018. A LGPD, inspirada pelo GDPR, estabelece princípios, direitos e deveres aplicáveis ao tratamento de dados pessoais, buscando garantir o respeito à autodeterminação informativa dos titulares. No contexto atual, o **consentimento do titular** destaca-se como uma das bases legais para o tratamento de dados, sendo compreendido como a concordância do titular de dados com o tratamento de seus dados, por meio de uma manifestação livre, informada e inequívoca, com uma finalidade específica [5].

No entanto, embora o consentimento seja exigido, inclusive em mídias sociais digitais, para o tratamento de dados dos usuários, frequentemente há uma assimetria de informações e de poder nas plataformas. Os usuários, embora teoricamente informados sobre as formas de tratamento de dados por meio das políticas de privacidade, muitas vezes não estão plenamente cientes de como ocorrem, de fato, a coleta, o uso e o compartilhamento de seus dados [2]. Isso ocorre principalmente devido ao volume e à complexidade dos termos e condições de uso e políticas de privacidade, que, além de conterem termos técnicos e complexos que muitos indivíduos não compreendem, exigiriam cerca de 400 horas por ano para serem lidas [6, 7], caso os usuários lessem todas as políticas às quais se deparam. No entanto, é justamente por meio desses documentos que o indivíduo compreende — ou deveria compreender — o que será feito com seus dados pessoais pelas plataformas

Além disso, fatores cognitivos e comportamentais levam os usuários a aceitarem termos que não compreendem totalmente, influenciados por vieses e pelo medo de exclusão social [8, 9]. Somado a isso, algumas interfaces exploram fatores cognitivos e utilizam dark patterns. Estes são práticas que manipulam e limitam a autonomia dos indivíduos, levando-os a decisões que não são do seu melhor interesse e podem causar prejuízos [10].

Assim, embora legislações como a LGPD e o GDPR exijam clareza, as plataformas frequentemente falham em tornar as informações acessíveis e equilibradas por meio de suas políticas, comprometendo a efetividade do consentimento [11]. Portanto, o consentimento pode não refletir uma escolha realmente informada, dificultando o exercício dos direitos dos usuários sobre seus dados pessoais.

Diante da complexidade e dos desafios do consentimento e dos fatores que o envolvem em mídias sociais digitais, a representação do conhecimento surge como uma construção conceitual para organizar, estruturar e formalizar as informações relevantes sobre esse domínio. A ontologia, enquanto artefato da representação do conhecimento, permite construir um modelo compreensível por humanos e máquinas, fornecendo uma compreensão clara e compartilhada dos elementos envolvidos no consentimento do titular dos dados [12].

Além de facilitar a interoperabilidade entre sistemas, a ontologia possibilita a atualização contínua frente às rápidas mudanças do ambiente digital, bem como a reutilização e integração com outras bases de conhecimento. Assim, a ontologia não apenas representa o conhecimento, mas o estrutura de modo a apoiar soluções para os desafios do consentimento nas mídias sociais digitais.

Diante do exposto, o presente artigo visa apresentar o desenvolvimento inicial, no âmbito de uma pesquisa de mestrado, de uma estrutura semântica ontológica voltada à representação do consentimento do titular de dados pessoais no contexto das mídias sociais digitais. Além de considerar os aspectos legais brasileiros relacionados à base legal do consentimento, a ontologia proposta também incorpora elementos que influenciam a obtenção desse consentimento. A proposta busca, assim, oferecer um modelo que una aspectos jurídicos e sociocomportamentais.

Este artigo está organizado em seções, conforme a seguir: a seção "Trabalhos Relacionados" apresenta outras ontologias de domínio relacionadas com o escopo do presente estudo; a seção "Metodologia Adotada" descreve detalhadamente as etapas e processos realizados, com vistas à reprodutibilidade científica; a seção "Resultados Parciais" explica a ontologia em desenvolvimento; por fim, a seção "Considerações Parciais e Trabalhos Futuros" finaliza com os últimos comentários sobre o artigo e com as possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

Em uma busca realizada nos repositórios de ontologias BioPortal[1], Ontology Lookup Service[2], TechnoPortal[3], FAIRsharing[4], OntoHub[5] e Linked Open Vocabularies[6], no repositório de código aberto GitHub[7], e nas bases de dados científicas Web of Science[8], Scopus[9] e SciELO[10], foram utilizadas as palavras-chave "*consent ontology*", "*personal data ontology*," "*protection ontology*", "*social media ontology*", "GDPR *ontology*", "LGPD *ontology*". A escolha desses termos visou abranger ontologias relacionadas ao consentimento, à proteção de dados pessoais e ao contexto das mídias sociais digitais, bem como identificar modelos alinhados ao GDPR, que serviu de referência para a legislação brasileira, e à LGPD. Como resultado, foram encontrados os seguintes trabalhos: *GConsent*[11], *GDPRov*[12], *Consent Ontology*, *GDPR Ontology*[13], *Ontology for the Protection of Personal Data* (OPPD), *PrOnto*, *OntoPriv, Ontology for Privacy Policies of OSNs* (OPPO)[14] e o *Data Privacy Vocabulary* (DPV)[15].

A *GConsent* é uma ontologia para representar o consentimento e seus diferentes estados em conformidade com o GDPR [13], sem um contexto de aplicação específica. Essa ontologia não modela aspectos como a finalidade ou a forma de tratamento dos dados. A *GDPRov* [14], a *GDPR Ontology*, a OPPD [15] e a *PrOnto* [16] também são ontologias desenvolvidas em conformidade com o GDPR. Enquanto a GDPRov representa o fluxo de dados, as demais têm um viés mais jurídico e regulatório. Todas essas quatro ontologias carecem de um contexto de uso definido.

A *Consent Ontology* [17] representa o consentimento com base na Lei de Proteção de Dados Pessoais nº 6.698 da Turquia. A *OntoPriv* [18], por sua vez, tem base na Lei Orgânica de Proteção de Dados Pessoais do Equador, e seu objetivo é assegurar conformidade regulatória. A OPPO modela práticas de tratamento de dados descritas nas políticas de privacidade de plataformas de mídias

[1] https://bioportal.bioontology.org
[2] https://www.ebi.ac.uk/ols4
[3] https://technoportal.hevs.ch
[4] https://fairsharing.org
[5] https://ontohub.org
[6] https://lov.linkeddata.es/dataset/lov
[7] https://github.com
[8] https://www.webofscience.com/wos
[9] https://www.scopus.com/home.uri
[10] https://www.scielo.br
[11] https://openscience.adaptcentre.ie/ontologies/GConsent/docs/ontology
[12] https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology
[13] https://github.com/ShahAJh/GDPR_Ontology_Project
[14] https://github.com/SanondaDattaGupta/OPPO-Ontology
[15] https://w3c.github.io/dpv/2.1/dpv

sociais digitais. Por fim, o DPV é um vocabulário que representa conceitos relacionados à privacidade e proteção de dados, derivados do GDPR.

Não foram encontradas ontologias que modelassem o consentimento de acordo com a LGPD, nem que estabelecessem conexões ou incorporassem aspectos relacionados a fatores que possam influenciar o consentimento fornecido pelo titular, no contexto das mídias sociais digitais.

3. Metodologia Adotada

Esta é uma **pesquisa qualitativa**, de caráter **descritivo** e de natureza **aplicada, com o objetivo geral de sistematizar** o conhecimento sobre o consentimento de indivíduos em mídias sociais digitais, visando apoiar soluções tecnológicas para a privacidade de dados pessoais, a conformidade normativa e a conscientização dos usuários. A metodologia escolhida para o desenvolvimento da ontologia foi a *OntoForInfoScience*, elaborada por Mendonça [19]. Essa metodologia, criada para ser utilizada por profissionais da Ciência da Informação, detalha de maneira simples e em linguagem acessível todas as etapas necessárias para o desenvolvimento de ontologias, superando barreiras encontradas pelos cientistas da informação, como jargões técnicos e questões filosóficas profundas [19].

Além de ser de fácil compreensão, a escolha pela *OntoForInfoScience* também foi motivada pelo seu potencial de contribuir para a disseminação do conhecimento sobre ontologias e do desenvolvimento das mesmas por pessoas que não são especialistas na área, principalmente entre os cientistas da informação. A *OntoForInfoScience* estrutura o processo em oito etapas, além de uma etapa preliminar (etapa 0).

3.1. Etapa 0 - Avaliação da Necessidade da Ontologia

Antes de iniciar o desenvolvimento da ontologia, a *OntoForInfoScience* prevê uma etapa preliminar para avaliar se há realmente uma necessidade para a criação da ontologia, ou se outro instrumento, como um tesauro ou uma taxonomia, seria suficiente para atingir o objetivo proposto [19].

Considerando que o objetivo é a representação do conhecimento de um determinado domínio, avaliou-se que outros instrumentos de recuperação da informação não seriam suficientes para atender à complexidade necessária. Planejou-se uma representação de aspectos do mundo real e a relação entre entidades deste, além de buscar possibilitar o uso em um modelo de mundo aberto, o que vai além das capacidades oferecidas por outros instrumentos.

A motivação também se sustenta na necessidade de explicitar formalmente os conceitos envolvidos, suas propriedades e suas inter-relações. Diferentemente de estruturas mais rígidas, a ontologia permite mais expressividade semântica, favorecendo a representação de significados mais complexos e contextuais. Propõe-se, ainda, que a ontologia seja extensível futuramente e reutilizável em outras ontologias, além de possibilitar inferência automática e aplicação na Web Semântica.

3.2. Etapa 1 - Especificação da Ontologia

A primeira etapa consiste no preenchimento da especificação da ontologia, que é um *template* contendo informações sobre o domínio, o propósito geral, a classe de usuários a que se destina, o uso pretendido, o tipo de ontologia, o grau de formalidade e a delimitação do escopo e as questões de competência [19].

O domínio abrange o **consentimento de indivíduos** em mídias sociais digitais, considerando os aspectos e **os requisitos da LGPD**. Parte-se da definição de consentimento prevista na LGPD, que o estabelece como uma das bases legais para o tratamento de dados pessoais. O propósito é oferecer uma estrutura formal que facilite a compreensão, a análise e a aplicação prática desses conceitos, promovendo a conformidade com a legislação brasileira, além de favorecer o entendimento de aspectos que podem influenciar ou comprometer o consentimento do titular.

Os usuários-alvo incluem pesquisadores das áreas de Ciência da Informação, Direito, Ciência da Computação e Psicologia, profissionais da área de privacidade e proteção de dados, desenvolvedores de sistemas, empresas, órgãos reguladores e o governo. O uso pretendido abrange tanto pesquisas científicas e acadêmicas quanto o desenvolvimento de sistemas e a interoperabilidade de dados com foco na padronização da representação do consentimento. A ontologia é classificada como uma ontologia de domínio, com grau de formalidade médio e estrutura semiformal.

Seu ponto de partida são entidades do mundo real descritas na Basic Formal Ontology (BFO) e o conceito de consentimento conforme definição na LGPD, delimitando-se à hipótese legal de obtenção do consentimento no contexto das mídias sociais digitais, sem abordar aspectos técnicos de implementação nem o tratamento de dados sensíveis ou de crianças e adolescentes. As questões de competência orientam o desenvolvimento da ontologia e incluem perguntas sobre tipos de consentimento, validade, revogação, agentes envolvidos, princípios legais aplicáveis e práticas que comprometem a autonomia do titular.

3.3. Etapa 2 - Aquisição e Extração do Conhecimento

A segunda etapa da metodologia consiste na adoção de métodos para aquisição e extração do conhecimento, com a seleção de fontes de informação [19]. Para essa etapa, foi realizada uma pesquisa bibliográfica e documental com o objetivo de compreender o domínio de estudo.

Por se tratar de uma pesquisa multidisciplinar, a pesquisa bibliográfica também foi conduzida em bases científicas de caráter multidisciplinar: Web of Science, Scopus e SciELO. A busca utilizou as seguintes *strings*: (“*personal data protection*” *AND* “*consent*”), (“*personal data protection*” *AND* “*digital social media*”), (“*personal data protection*” *AND* “*social media*”), (“*consent*” *AND* “*digital social media*”), (“*consent*” *AND* “*social media*”), (“*personal data protection*” *AND* “*consent*” *AND* “*digital social media*”), ("*personal data protection*" *AND* “*consent” AND* “*social media”*), (“proteção de dados pessoais” *OR* “LGPD”), ("*General Data Protection Regulation*" *OR* “GDPR” *OR* "Regulamento Geral sobre a Proteção de Dados" *OR* “RGPD”). Foram considerados artigos em inglês e português, de acesso aberto. Após a exclusão dos duplicados e o *screening*[16] dos artigos, foram selecionados 55 documentos para serem estudados, com o objetivo de aprofundar a compreensão do domínio.

A pesquisa documental foi realizada por meio da consulta a fontes normativas brasileiras, incluindo a Constituição Federal, a LGPD, o Código Civil, a Lei de Acesso à Informação, o Marco Civil da Internet e o Código de Defesa do Consumidor, além de resoluções e outros documentos emitidos pela Autoridade Nacional de Proteção de Dados. Também foram analisados documentos de organizações não governamentais, como o Comitê Gestor da Internet (CGI.br) e a Organização Internacional de Padronização (ISO). Adicionalmente, foram consultadas fontes de organizações internacionais, como a Organização para a Cooperação e Desenvolvimento Econômico (OCDE).

Visando ao aprofundamento do domínio em estudo e ao entendimento dos aspectos legais das regulamentações que envolvem a área, o processo de extração dos termos candidatos a classes foi realizado de forma manual.

3.4. Etapa 3 - Conceitualização

A próxima etapa refere-se à conceitualização, ou seja, à elaboração da tabela de conceitos e propriedades, do dicionário de verbos e dos modelos conceituais gráficos [19]. Durante essa etapa, foram extraídos 156 termos das fontes de informação, acompanhados de seus respectivos conceitos, compilados em uma planilha eletrônica.

A tabela foi revisada, com a exclusão de termos repetidos e daqueles considerados não pertinentes para uso futuro como classes da ontologia. Ao final, restaram 130 termos. Para cada termo, foram atribuídos, em colunas da planilha eletrônica, os seguintes metadados: ID**, *class* (EN),**

[16] Processo de triagem para selecionar e avaliar os estudos que serão incluídos com base nos títulos, resumos e palavras-chave.

label **(EN),** ***label*** **(PT),** ***alt label*****,** ***comment or definition, is defined by, parent, has subclasses, type, object properties*** **e** ***semi-formal definition*****.**

A coluna *object properties* foi construída a partir dos próprios conceitos, como, por exemplo, "[*consent is*] *deliberately* ***granted by*** *the data subjects*". Essas relações foram organizadas em uma nova planilha, com o objetivo de criar o dicionário de verbos. Para cada relação, foram atribuídos os seguintes metadados: *object property*, *source ontology*, *synonyms*, *inverse property*, *definition*, *usage example* e *characteristics*. Estas últimas referem-se às características das propriedades, tais como simétrica, funcional ou reflexiva.

Os modelos conceituais gráficos, embora façam parte dessa etapa da metodologia, não foram realizados nesse momento, visto que a representação gráfica pode ser criada após a linguagem lógica estar completa e exportada.

3.5. Etapa 4 – Fundamentação Ontológica

A quarta etapa consiste em pesquisar, escolher e aplicar ontologias de fundamentação no desenvolvimento da ontologia. Assim, após uma análise inicial, a BFO foi selecionada como ontologia de fundamentação por oferecer uma estrutura que distingue as entidades entre continuantes (como agentes e dados) e ocorrentes (como os atos de consentimento). **No entanto, embora alguns elementos da BFO tenham sido importados para testes iniciais, a sua integração definitiva e o alinhamento completo ainda não foram concluídos.** Essa incorporação será realizada nas fases subsequentes do desenvolvimento.

3.6. Etapa 5 – Formalização da Ontologia

A quinta etapa consiste na representação formal de todo o conhecimento previamente elaborado, utilizando uma linguagem descritiva ou lógica que permita a interpretação pelas máquinas, favorecendo inferências e a interoperabilidade. Assim, o conhecimento conceitual passa a um nível ontológico-formal [19].

Esta etapa inclui também a pesquisa e a identificação de classes equivalentes em outras ontologias de domínio, visando a integração e a reutilização do modelo desenvolvido. Para isso, foram consultados repositórios e portais de ontologias, como BioPortal, Linked Open Vocabularies, Ontology Lookup Service, OntoHub e GitHub. A partir dessa análise, as classes consideradas compatíveis com o escopo da ontologia foram adicionadas na coluna "termos similares" da tabela de conceitos e propriedades.

Para a etapa de formalização, foi utilizado o software Protégé, versão 5.6.5. O processo teve início com a inserção dos termos, dados e metadados da tabela de conceitos e propriedades, bem como as relações identificadas no dicionário de verbos. Também foram importadas as classes equivalentes de outras ontologias de domínio, previamente identificadas.

A última parte consistiu na adição de axiomas, quando pertinentes, com a especificação de restrições e relações entre as classes e propriedades. Por exemplo, definiu-se que a classe ***ConsentGiven***, subclasse de ***gc:ConsentStatusValidForProcessing,*** é equivalente à interseção simultânea das classes ***FreeConsent***, ***UnequivocalConsent*** e ***InformedConsent***. Além disso, o ***ConsentGiven*** deve ser concedido (por meio da propriedade ***isGrantedBy***) por exatamente um indivíduo pertencente à classe ***DataSubject***. Em outras palavras, o tratamento de dados é permitido quando o consentimento dado pelo titular é livre, informado e inequívoco.

3.7. Etapa 6 – Avaliação da Ontologia

A sexta etapa corresponde à avaliação da ontologia, composta pelas fases de validação - que avalia a adequação da ontologia em relação ao domínio representado - e de verificação, que avalia a consistência ontológica, ambas realizadas com base em critérios avaliativos [19].

Conforme recomendado na própria metodologia *OntoForInfoScience*, foram utilizados os *reasoners* HermiT e Pellet, disponíveis no próprio Protégé, para a validação lógica e para a verificação de problemas de imprecisão e inconsistência nas relações semânticas. As inconsistências apontadas pelos *reasoners* foram analisadas, revisadas e corrigidas. Além disso, a ontologia passou por uma revisão com consulta a outros pesquisadores, com o objetivo de assegurar a qualidade conceitual e a aderência aos requisitos do domínio.

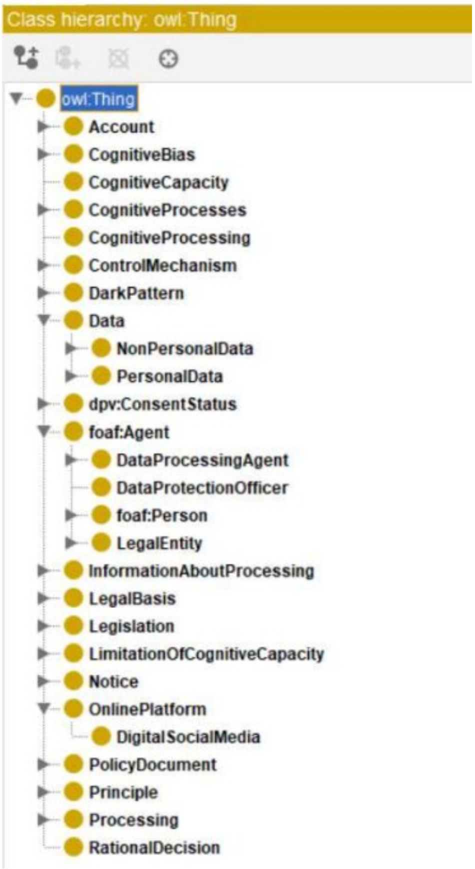
3.8. Etapas 7 e 8 - Documentação e Disponibilização da Ontologia

As duas últimas etapas devem ser realizadas após a conclusão efetiva de todas as etapas anteriores, consistindo na documentação formal, em linguagem natural, de todas as informações contidas na ontologia e, por fim, na disponibilização da ontologia em linguagem lógica, apresentada em meio eletrônico [19].

O desenvolvimento da proposta apresentada neste artigo ocorreu entre janeiro e julho de 2025. Até a data limite para a submissão deste artigo, a pesquisa encontrava-se em fase final de desenvolvimento. Portanto, a documentação final ainda está em fase de elaboração e a ontologia ainda não foi disponibilizada publicamente. Contudo, planeja-se disponibilizá-la em acesso aberto, em formato *Web Ontology Language* (OWL), por meio de um repositório de dados.

4. Resultados Parciais

A pesquisa bibliográfica e documental resultou na extração e na formalização de 130 classes para compor a base conceitual da ontologia. A organização hierárquica dessas classes (figura 1)estrutura-se em eixos principais para a representação do conhecimento, tais como: o agente e seus processos cognitivos (ex: ***foaf:Agent***, ***CognitiveBias***, ***RationalDecision***), as plataformas de mídias sociais e seus mecanismos (ex: ***OnlinePlatform***, ***ControlMechanism***, ***DarkPattern***), os instrumentos normativos e informativos (ex: ***Legislation***, ***LegalBasis***, ***PolicyDocument***, ***Notice***); e o tratamento de dados (ex: ***Data***, ***Processing***, ***dpv:ConsentStatus***). Essa estruturação visa representar de forma conectada os diversos elementos que permeiam o consentimento do titular.



- owl:Thing
  - Account
  - CognitiveBias
  - CognitiveCapacity
  - CognitiveProcesses
  - CognitiveProcessing
  - ControlMechanism
  - DarkPattern
  - Data
    - NonPersonalData
    - PersonalData
  - dpv:ConsentStatus
  - foaf:Agent
    - DataProcessingAgent
    - DataProtectionOfficer
    - foaf:Person
    - LegalEntity
  - InformationAboutProcessing
  - LegalBasis
  - Legislation
  - LimitationOfCognitiveCapacity
  - Notice
  - OnlinePlatform
    - DigitalSocialMedia
  - PolicyDocument
  - Principle
  - Processing
  - RationalDecision

Figura 1: Disposição hierárquica de parte das classes da ontologia mostrada na aba class hierarchy no Protégé.

4.1. O Consentimento

O conceito central da ontologia é a classe ***Consent*** (figura 2) que representa a concordância do titular para o tratamento de seus dados pessoais [5], e figura como uma subclasse de ***LegalBasisUnderLGPD***. Para que o consentimento seja considerado válido, ele precisa ser livre, informado e inequívoco [5]. A intersecção dessas qualificadoras (***FreeConsent***, ***InformedConsent*** e ***UnequivocalConsent***) resulta na classe ***ConsentGiven***, que é concedido por exatamente um ***DataSubject***.

Utilizando classes importadas do DPV e da GConsent, também são modelados os diferentes status que o consentimento pode assumir. A classe ***dpv:ConsentStatus*** é especializada em ***gc:ConsentStatusValidForProcessing*** e ***gc:ConsentStatusInvalidForProcessing***. Enquanto o tratamento de dados (***Processing***) só pode ser realizado sob uma única condição, um ***ConsentGiven***, o tratamento não pode ser efetuado nas seguintes condições [5]: ***ConsentRefused***, quando o titular não aceita o tratamento; ***ConsentWithdrawn***, quando o titular retira o consentimento previamente dado; ***NullConsent***, quando as informações fornecidas ao titular sobre o tratamento são enganosas ou não transparentes; e ***ConsentVitiated***, quando o consentimento é obtido por meio de defeitos do negócio jurídico (coerção, dolo, erro, estado de perigo).

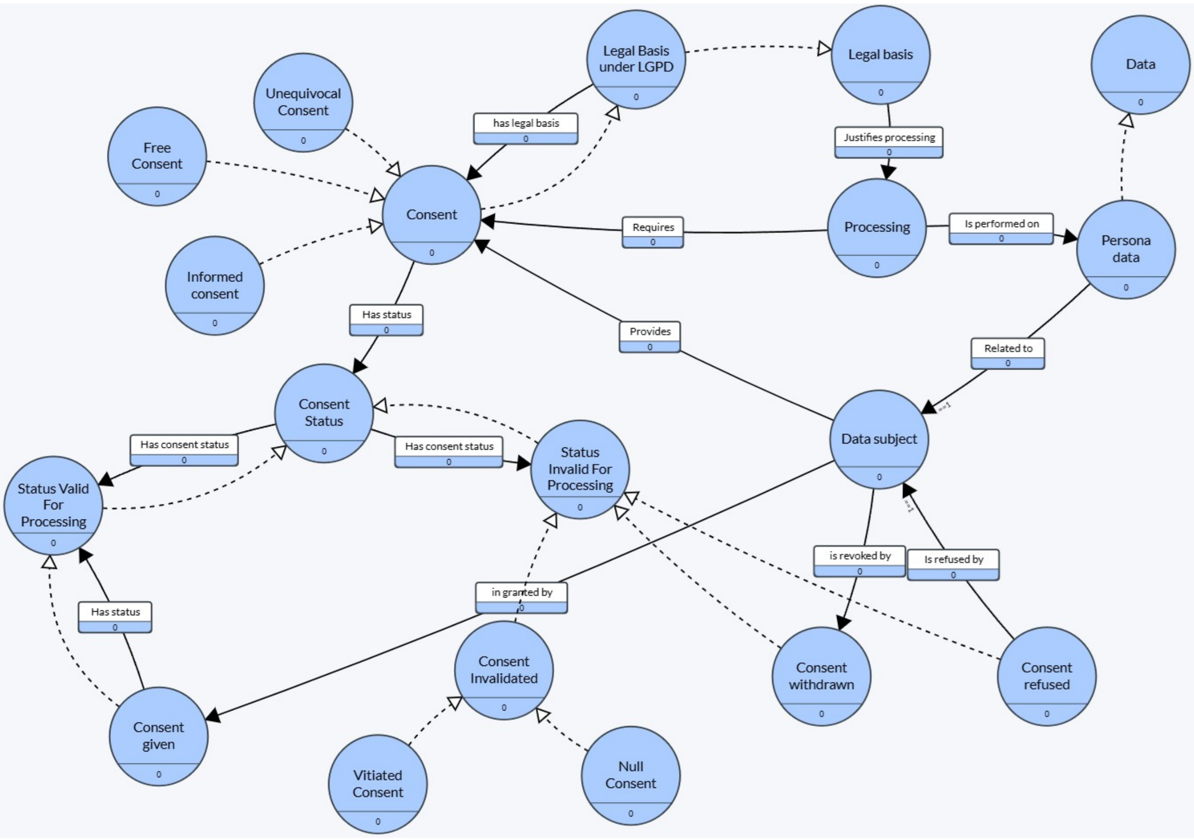


Figura 2: Diagrama que modela o consentimento como base legal para o tratamento de dados pessoais, destacando seus tipos, estados e relações com o titular dos dados segundo a LGPD.

4.2. O Titular de Dados e seus Aspectos Cognitivos

O consentimento é articulado em torno de uma figura central: o ***DataSubject***, que é a pessoa a quem se referem os dados pessoais que são objeto de tratamento [5]. Nesse contexto, o ***DataSubject*** é uma subclasse de ***NaturalPerson***, que, por sua vez, é subclasse de ***foaf:Person*** e ***foaf:Agent***. No contexto específico das mídias sociais digitais, o usuário de mídias sociais digitais (***DigitalSocialMediaUser***) é uma especificação de ***DataSubject***.

A ontologia dedica atenção aos aspectos cognitivos que podem influenciar a capacidade do titular de dados de fornecer um consentimento verdadeiramente livre, informado e inequívoco (figura 3). A classe ***CognitiveCapacity*** representa a capacidade cognitiva e, também, os limites ao processamento cognitivo, que são processos mentais que a mente utiliza para processar e organizar informações [20]. As limitações dessa capacidade são modeladas pela classe ***LimitationOfCognitiveCapacity***, que pode tornar o processo de tomada de decisão (***DecisionMakingProcess***) mais difícil e influenciar uma decisão racional (***RationalDecision***).

Adicionalmente, a ontologia modela a classe ***CognitiveBias***, que são distorções cognitivas que podem desviar a percepção, o julgamento e a tomada de decisão [21]. As subclasses de vieses cognitivos necessitam de expansão. No momento, englobam somente as classes ***DefaultBias***, ***FreeBias*** e ***PresentBias***.

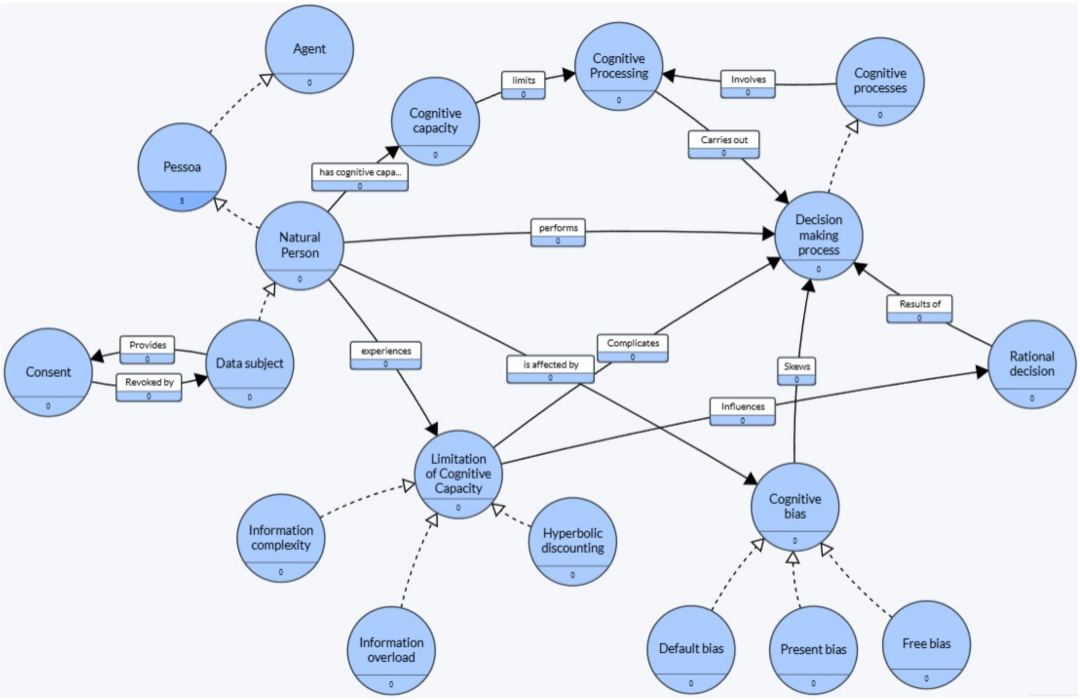


Figura 3: Diagrama que modela as limitações cognitivas e vieses que afetam o processo decisório da pessoa natural na concessão do consentimento.

4.3. Mídias Sociais Digitais e seus Mecanismos

No escopo deste trabalho, a interação do titular para fins de consentimento ocorre em mídias sociais digitais (***DigitalSocialMedia***), que são uma subclasse de ***OnlinePlatform***. O ***DigitalSocialMediaUser*** está ligado a essas plataformas por meio da posse de uma conta (***DigitalSocialMediaAccount***) nelas.

A ontologia também representa os mecanismos presentes nas plataformas que gerenciam as preferências de consentimento (figura 4). A classe ***ConsentMechanism*** tem a subclasse ***ConsentManagement***, que são sistemas ou processos que permitem aos usuários determinarem quais informações eles permitem que sejam acessadas [22]. Um exemplo é o mecanismo ***OptIn***, que exige uma ação afirmativa do usuário para a concessão do consentimento [23].

Um foco particular da ontologia é a representação da classe ***DarkPattern***, definido como interfaces implementadas nas plataformas que tentam influenciar os usuários a tomarem decisões não intencionais, involuntárias e potencialmente prejudiciais em favor dos interesses das plataformas [24]. São detalhados os seguintes tipos de *dark patterns*: ***Fickle***, caracterizado por uma interface inconsistente e não clara**;** ***LeftInTheDark***, quando a interface é projetada para ocultar informações; ***Obstructing,*** quando as interfaces dificultam o gerenciamento de dados; ***Overloading***, quando o usuário se depara com uma sobrecarga de informações; ***Skipping***, caracterizado pelo *design* projetado de forma que os usuários esqueçam sobre a proteção de dados; e ***Stirring,*** que afeta a escolha do usuário ao apelar para as suas emoções [24].

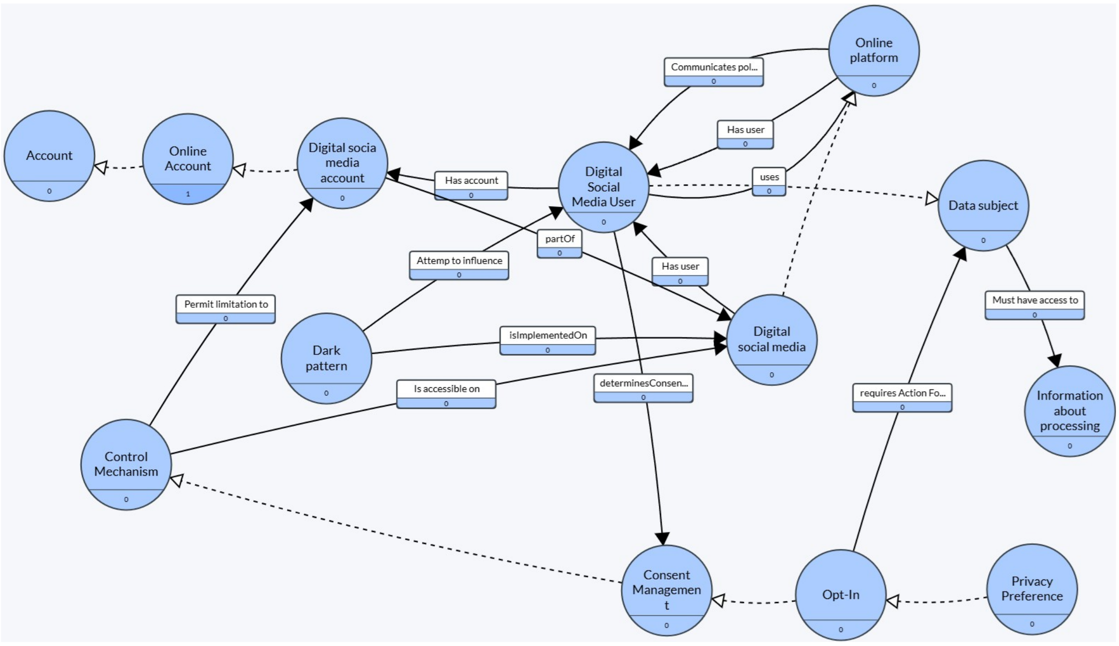


Figura 4: Diagrama representando as relações entre usuários, plataformas e mecanismos de consentimento em mídias sociais digitais.

4.4. Arcabouço Legal-Normativo e Informativo

A classe ***Legislation*** especializa-se em ***BrazilianLegislation***, que, por sua vez, tem como instância a LGPD e o Código Civil Brasileiro. O Código Civil define os vícios do negócio jurídico, como erro, dolo ou coação, que podem invalidar o consentimento, caracterizando o vício de consentimento, um tipo de consentimento inválido para o tratamento de dados [5]. Além disso, o Código também trata da capacidade jurídica necessária para que o negócio jurídico seja válido, exigindo a livre manifestação da vontade. A LGPD, por sua vez, é referência para estabelecer as hipóteses legais para o tratamento de dados pessoais, representadas pela classe ***LegalBasisUnderLGPD***, que é uma especialização de ***LegalBasis***. Dentro do domínio da ontologia, a classe ***Consent*** é modelada como uma das hipóteses legais prevista na LGPD (figura 5).

Os princípios que o tratamento de dados pessoais deve observar são representados como subclasses de ***PrincipleUnderLGPD,*** que por sua vez é uma especialização de ***Principle.*** A ontologia representa os seguintes princípios: ***Purpose***, ***Adequacy***, ***Necessity***, ***FreeAccess***, ***DataQuality***, ***Transparency***, ***Security***, ***Prevention***, ***NonDiscrimination***, ***Accountability*** e ***Liability*** [5]. Além disso, também são representados os diferentes tipos de operações de tratamento de dados.

Também são representados os agentes no ciclo de tratamento de dados pessoais, subclasses da classe ***foaf:Agent***. Os agentes de tratamento são o ***Controller,*** que é o responsável pelas decisões sobre o tratamento, e o ***Processor,*** que efetua o tratamento em nome do controlador [5]. O ***DataProtectionOfficer*** atua como canal de comunicação entre o ***Controller***, o ***DataSubject*** e a ***NationalAuthority*** [5], representada pela Autoridade Nacional de Proteção de Dados.

A ontologia também estrutura as informações que devem ser fornecidas ao titular dos dados para a obtenção de um consentimento válido. As informações incluem a duração e a forma de tratamento, as informações do controlador, as responsabilidades do agente e os direitos do titular. Os direitos dos titulares são agrupados sob a classe ***DataSubjectRight.***

Por fim, é representada a classe ***PolicyDocument***, que corresponde aos documentos disponibilizados pelas mídias sociais digitais. Essa classe é subdividida em ***PrivacyPolicy*** e ***TermsAndConditions***. A ontologia estabelece que uma mídia social digital deve possuir tais documentos, e que estes devem ser acessíveis aos usuários. A ***PrivacyPolicy*** descreve as operações de tratamento de dados e deve fornecer todas as informações sobre o tratamento ao ***DataSubject***, aspecto essencial para a validade do consentimento.

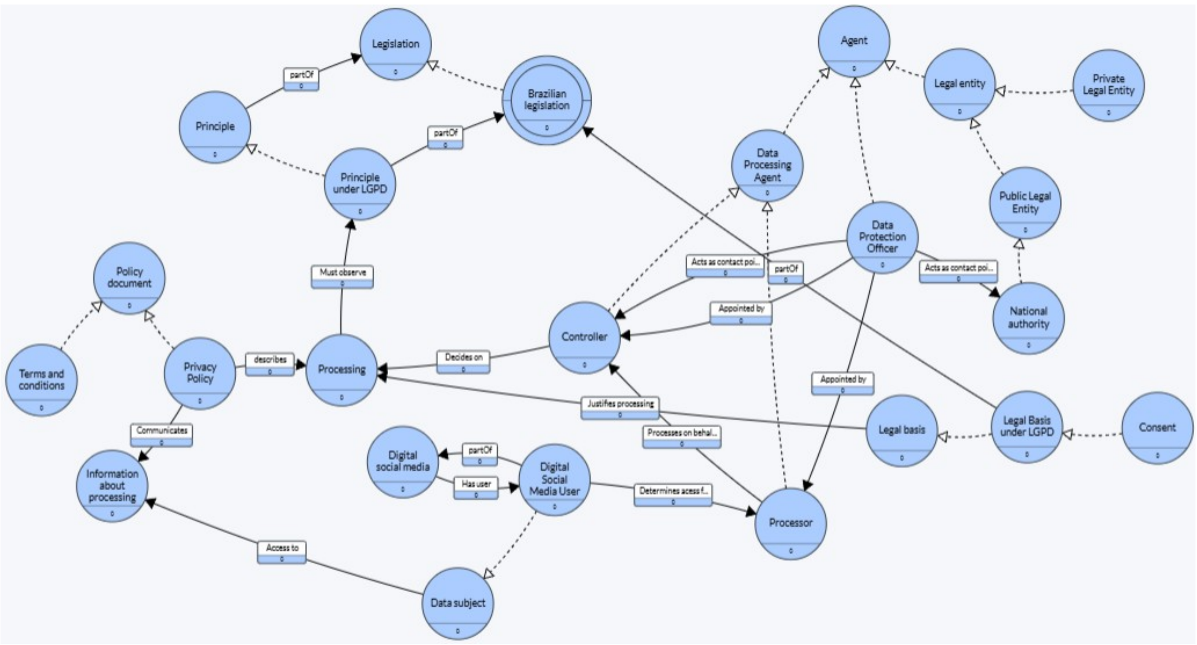


Figura 5: Diagrama representando agentes e fundamentos envolvidos no tratamento de dados pessoais em mídias sociais digitais, conforme a LGPD.

5. Considerações Finais e Trabalhos Futuros

O presente artigo apresentou resultados parciais do desenvolvimento de uma ontologia de domínio para a representação do conhecimento sobre o consentimento do titular de dados pessoais, no contexto de mídias sociais digitais. A partir da análise de documentos formais e científicos, foi possível identificar e estruturar os principais conceitos e relações do domínio, considerando tanto aspectos legais quanto sociocomportamentais. A proposta busca refletir a complexidade do tema, incluindo fatores que afetam a validade do consentimento, como a presença de *dark patterns* e os limites da racionalidade dos usuários. A ontologia está sendo desenvolvida com base na metodologia *OntoForInfoScience* e reutiliza classes das ontologias DPV, *GConsent*, APAONTO[17] e FOAF[18]. Ao todo, a ontologia é composta por 130 classes, 52 propriedades de objeto, 1.234 axiomas, sendo que 334 são axiomas lógicos e 211 são declarações de axiomas. Visando ampliar sua reutilização, os termos estão anotados no *rdfs:label* em versões bilíngues inglês e português.

Espera-se que essa ontologia possa auxiliar pesquisadores, desenvolvedores de sistemas, profissionais da área jurídica, representantes do governo e de empresas no entendimento estruturado dos elementos que envolvem o consentimento nas mídias sociais digitais, contribuindo para o desenvolvimento de soluções alinhadas às exigências legais e éticas. Além disso, espera-se que a ontologia beneficie os próprios usuários das plataformas, possibilitando a compreensão de seus direitos e dos mecanismos que muitas vezes estão ocultos em estruturas manipulativas de interface, favorecendo, assim, a literacia digital.

Como trabalhos futuros, pretende-se realizar a integração com a ontologia de fundamentação BFO, bem como validar e avaliar a ontologia. Após devidamente finalizada, planeja-se disponibilizá-la em formato OWL em um repositório de acesso aberto, incentivando o reuso. Também espera-se, futuramente, expandir a ontologia para contemplar outros aspectos da proteção de dados pessoais, bem como aprofundar os aspectos cognitivos que influenciam o comportamento dos usuários nas mídias sociais digitais, como o compartilhamento de informações e a tomada de decisão em contextos digitais.

[17] https://bioportal.bioontology.org/ontologies/APAONTO
[18] http://xmlns.com/foaf/spec

Agradecimentos

Agradecemos ao Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR) pelo suporte à realização deste trabalho. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), por meio de recursos do PROAP, Edital nº 28/2025 – SA/PPGGI.

Declaração de IA Generativa

Durante a preparação deste trabalho, os autores utilizaram o GPT-4 para: Verificação gramatical e ortográfica. Após a utilização da ferramenta, os autores revisaram e editaram o conteúdo conforme necessário e assumem responsabilidade total pelo conteúdo da publicação.

Referências

[1] D. S. Paulichi, V. S. G. Cardin, Dinâmicas da sociedade informacional contemporânea: análise da captação de dados em plataformas digitais e suas implicações socioculturais, Pensar 29 (2024) 1–14. doi:10.5020/2317-2150.2024.14816.

[2] N. F. D. M. Maciel, A globalização das plataformas digitais: uma análise sobre a necessidade de regulamentação dessa ferramenta, Rev. Foco 16 (2023) e3092. doi:10.54751/revistafoco.v16n10-155.

[3] X. Ye, Y. Yan, J. Li, B. Jiang, Privacy and personal data risk governance for generative artificial intelligence: A Chinese perspective, Telecommunications Policy 48 (2024) 102851. doi:10.1016/j.telpol.2024.102851.

[4] European Parliament and Council. Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, vol. L119, pp. 1–88, May 2016. URL: https://gdpr-info.eu.

[5] Brazil. Law No. 13,709, August 14, 2018 — General Data Protection Law (LGPD). Official Gazette of the Union, Brasília, 2018. URL: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

[6] V. Belcheva, T. Ermakova, B. Fabian, Understanding Website Privacy Policies—A Longitudinal Analysis Using Natural Language Processing, Information 14 (2023) 622. doi:10.3390/info14110622.

[7] X. Ding, H. Huang, For whom is privacy policy written? A new understanding of privacy policies, Computer Law & Security Review 55 (2024) 106072. doi:10.1016/j.clsr.2024.106072.

[8] L. Wolmarans, A. Voorhoeve, What Makes Personal Data Processing by Social Networking Services Permissible?, Canadian Journal of Philosophy 52 (2022) 93–108. doi:10.1017/can.2022.4.

[9] L. N. Lugati, J. E. D. Almeida, Da evolução das legislações sobre proteção de dados: a necessidade de reavaliação do papel do consentimento como garantidor da autodeterminação informativa, RD 12 (2020) 01–33. doi:10.32361/2020120210597.

[10] OECD. Dark commercial patterns. OECD Digital Economy Papers, no. 336, OECD Publishing, Paris, 2022. URL: https://doi.org/10.1787/44f5e846-en.

[11] D. Green, Strategic Indeterminacy and Online Privacy Policies: (Un)informed Consent and the General Data Protection Regulation, International Journal for the Semiotics of Law 38 (2025) 701–729. doi:10.1007/s11196-024-10132-4.

[12] M. B. Almeida, Ontologia em Ciência da Informação: Teoria e Método, vol. 1, Coleção Representação do Conhecimento em Ciência da Informação, Editora CRV, Curitiba, Brazil, 2020. ISBN 978-65-5578-679-8. doi:10.24824/978655578679.8.

[13] H. J. Pandit, C. Debruyne, D. O'Sullivan, D. Lewis, GConsent – A Consent Ontology Based on the GDPR, in: P. Hitzler, M. Fernández, K. Janowicz (Eds.), The Semantic Web: Proceedings of the 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, volume

11503 of Lecture Notes in Computer Science, Springer, Cham, 2019, pp. 270–282. doi:10.1007/978-3-030-21348-0_18.

[14] H. J. Pandit, D. Lewis, Modelling Provenance for GDPR Compliance using Linked Open Data Vocabularies, in: F. Scharffe, F. Schneider (Eds.), Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web – Policy and Technology (PrivOn 2017), co-located with ISWC 2017, volume 1951 of CEUR Workshop Proceedings, CEUR-WS.org, Vienna, Austria, 2017. URL: https://ceur-ws.org/Vol-1951/PrivOn2017_paper_6.pdf.

[15] M. El Ghosh, H. Abdulrab, Capturing the Basics of the GDPR in a Well-Founded Legal Domain Modular Ontology, Frontiers in Artificial Intelligence and Applications 344 (2021) 144–158. doi:10.3233/FAIA210378.

[16] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo, PrOnto: Privacy Ontology for Legal Reasoning, in: A. Kő, E. Francesconi (Eds.), Electronic Government and the Information Systems Perspective. EGOVIS 2018, Springer, Cham, 2018, pp. 101–115. doi:10.1007/978-3-319-98349-3_11.

[17] E. Olca, O. Can, DICON: A Domain-Independent Consent Management for Personal Data Protection, IEEE Access 10 (2022) 95479–95497. doi:10.1109/ACCESS.2022.3204970.

[18] G. Suntaxi, K. Ojeda, F. Rodríguez, OntoPriv: Enhancing Understanding and Compliance in Privacy Legislation via Legal Ontologies, in: Proceedings of the 2024 Latin American Computer Conference (CLEI), IEEE, Buenos Aires, Argentina, 2024, pp. 1–10. doi:10.1109/CLEI64178.2024.10700326.

[19] F. M. Mendonça, OntoForInfoScience: metodologia para construção de ontologias pelos cientistas da informação - uma aplicação prática no desenvolvimento da ontologia sobre componentes do sangue humano (Hemonto), Ph.D. thesis, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil, 2015. URI: http://hdl.handle.net/1843/BUBD-A35H3K.

[20] D. Krch, Cognitive processing, in J. S. Kreutzer, J. DeLuca, and B. Caplan (Eds.), Encyclopedia of Clinical Neuropsychology, New York, NY: Springer, 2011, p. 627, doi: 10.1007/978-0-387-79948-3_1443. URL: https://link.springer.com/referenceworkentry/10.1007/978-0-387-79948-3_1443.

[21] APAONTO – Ontologia sobre aspectos psicológicos e jurídicos do consentimento. NCBO BioPortal, 2024. URL: https://bioportal.bioontology.org/ontologies/APAONTO.

[22] N. Almeida, M. Silva, A Blockchain-Based Hybrid Architecture for Auditable Consent Management, IEEE Transactions on Industrial Informatics 19 (2023) 2654–2662. doi:10.1109/TII.2023.10604861.

[23] Associação Brasileira de Normas Técnicas – ABNT. NBR ISO/IEC 29100:2024 – Tecnologia da informação – Técnicas de segurança – Arquitetura de privacidade, Rio de Janeiro: ABNT, 2024.

[24] European Data Protection Board – EDPB. Guidelines 03/2022 on deceptive design patterns in social media platform interfaces: how to recognise and avoid them, versão final, Bruxelas, 24 fev. 2023. URL: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-032022-deceptive-design-patterns-social-media_en.

LLM Assisted Vocabulary Harmonization

Maria Claudia Reis Cavalcanti[1], Samir de Oliveira Ramos[1], Ronaldo Ribeiro Goldschmidt[1], Wallace Anacleto Pinheiro[1], Alexandra Miguel Raibolt da Silva[1], Alex Garcia[1], Bernardo Alkmim[2], Robinson Callou[2], Edward Hermann Haeusler[2], Cecília de Azevedo Castro César[3], Ferrucio de Franco Rosa[3] and José Maria Parente de Oliveira[3,†]

[1] *Military Institute of Engineering (IME), Praça Gen. Tiburcio, 80, Rio de Janeiro - RJ, 22290-270, Brazil*
[2] *Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225, Rio de Janeiro -RJ, 22451-900, Brazil*
[3] *Aeronautics Institute of Technology (ITA), Praça Marechal Eduardo Gomes, 50, São José dos Campos - SP, 12228-900, Brazil*

Abstract

One of the early challenges in ontology creation is developing a standardized vocabulary with clear definitions, especially when integrating concepts from various and often conflicting sources. We propose the LLM-Assisted Vocabulary Harmonization (LAVOHA) method, which leverages large language models (LLMs) to systematically analyze and reconcile concept definitions in natural language. Our approach is demonstrated in the cybersecurity domain, where we harmonized definitions from multiple established cybersecurity vocabularies. In a case study, the LAVOHA definitions were evaluated against a human consensus using criteria such as clarity, completeness, and alignment with expert understanding. The results indicate that LAVOHA produces definitions that are more consistent and comprehensive than those generated by an LLM without harmonization guidance. These findings suggest that LAVOHA can significantly enhance the quality and interoperability of ontology vocabularies in complex domains.

Keywords

LLM, Ontology, Vocabulary Harmonization, BM25, RAG

1. Introduction

According to the literature on ontology engineering [1, 2, 3], when formalizing an ontology, building or reusing a glossary of terms is demanded. However, existing glossaries may have problematic definitions due to ambiguous and conflicting terminology. In the cybersecurity incident response domain, many official standardization documents point to conflicting definitions of terms. Identifying the best definition of terms in this context is hard, time consuming, and requires great human effort.

Upon identifying the definitions for the same concept, merging them is a complex task fraught with several key problems, such as inconsistencies, name conflicts, and redundant hierarchies, to name a few. In summary, merging different descriptions of the same concept risks either loss of detail or overgeneralization. The process may fail to distinguish between closely related but distinct concepts, collapsing them into a single broad concept and losing important nuances. In contrast, merging truly identical concepts could be neglected, leading to unnecessary duplication.

In our literature mapping, we identified the need for agile approaches aimed at helping humans

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

[†] These authors contributed equally.

yoko@ime.eb.br (M. C. R. Cavalcanti); samir.ramos@ime.eb.br (S. d. O. Ramos); ronaldo.rgold@ime.eb.br (R. R. Goldschmidt); wallaceapinheiro@gmail.com (W. A. Pinheiro); raibolt@ime.eb.br (A. M. R. d. Silva); garcia@ime.eb.br (A. Garcia); balkmim@inf.puc-rio.br (B. Alkmim); robinson.rcmbf@gmail.com (R. Callou); hermann@inf.puc-rio.br (E. H. Haeusler); cecilia@ita.br (C. d. A. C. César); ferrucio@ita.br (F. d. F. Rosa); parente@ita.br (J. M. P. d. Oliveira)

0000-0003-4965-9941 (M. C. R. Cavalcanti); 0009-0004-6055-7182 (S. d. O. Ramos); 0000-0003-1688-0586 (R. R. Goldschmidt); 0000-0001-7076-8785 (W. A. Pinheiro); 0000-0002-8982-596X (A. M. R. d. Silva); 0000-0002-2649-1106 (A. Garcia); 0000-0002-6927-6174 (B. Alkmim); 0009-0005-3941-3309 (R. Callou); 0000-0002-4999-7476 (E. H. Haeusler); 0000-0002-0332-3759 (C. d. A. C. César); 0000-0001-9504-496X (F. d. F. Rosa); 0000-0002-7803-1718 (J. M. P. d. Oliveira)

agree on consensual definitions. Natural language processing techniques have been used in ontology engineering and, when effectively fine-tuned, LLMs might work as suitable assistants for ontology construction [4].

We propose the LLM Assisted Vocabulary Harmonization (LAVOHA) method to harmonize definitions from various conflicting sources (e.g., vocabularies and glossaries). Documents are segmented into smaller chunks and transformed into vector embeddings for efficient storage. When a query is received, the system retrieves the most relevant chunks using a similarity function and provides them as a context.

When building an ontology from natural language definitions, engaging in full alignment with existing ontologies is typically premature. At this stage, the focus should be on clarifying concepts, normalizing terminology, and establishing initial structures, tasks that require semantic flexibility and lack the grounding needed for reliable correspondence with formal models. Premature ontology alignment risks distorting the intended meaning or introducing misinterpretations.

To mitigate these risks, the early phase must prioritize vocabulary construction: identifying and standardizing domain-relevant terms to ensure internal coherence. Although this process has not yet involved mapping to external ontologies, it lays the foundation for future alignment by creating a stable semantic base. In this sense, vocabulary construction acts as a form of pre-alignment, shaping definitions and relationships that will later facilitate integration.

For these reasons, this work does not address the ontology alignment itself but the prior step of harmonizing vocabulary definitions. Alignment becomes feasible only once the core vocabulary and conceptual structures have stabilized.

The remainder of this paper is organized as follows. Section 2 provides theoretical foundations; Section 3 presents a synthesis of the literature review, focusing on related work; Section 4 introduces the LLM-Assisted Vocabulary Harmonization method; Section 5 describes a case study on applying the proposed method and discusses the results; and Section 6 presents the conclusions.

2. Foundations

2.1. Ontology Building Methodologies

Most of the methodologies for ontology engineering found in the literature [1, 2, 3] include sub-processes, ranging from requirements elicitation to testing. Usually, in the requirements elicitation sub-process, they gather existing vocabularies and other related standard and reference documents. Figure 1 shows a variation of the main sub-processes (white boxes) of the SABIO methodology [1] in BPMN notation. The conceptualization/formalization sub-process embeds a vocabulary construction task, in which the ontologist must define the terms that can be covered by the ontology. Note that in Figure 1 this task is reified (gray box) and represented as a sub-process. We treated this task as a sub-process because its complexity is greater than it first seemed

According to some authors [5] [6], building a vocabulary, i.e. a list of terms and their corresponding definitions, is not a trivial task and may affect the quality and consistency of the ontology under construction. Many challenges emerge and must be solved with the support of domain experts. Some of the main issues that should be addressed are [5]: (i) *Inconsistencies* – there are plenty of terminological resources, and different definitions coming from these resources can be conflicting and lead to logical contradictions; (ii) *Homonyms* – two different concepts might share the same name but refer to different things, resulting in either unintended duplication and ambiguity, and should be distinguished by using different terms; (iii) *Synonyms* - multiple terms that refer to the same concept may create confusion, thus a preferred term must be chosen; (iv) *Recursive or Circular Definitions* - definitions that define a term in terms of itself, or in terms of another that references it, can create logical inconsistencies, and should be avoided; (v) *Redundant Hierarchies* – merging can introduce multiple paths between concepts or duplicate subclass relationships, cluttering the structure and making maintenance difficult.

Based on these issues, a set of tasks should be planned, such as reconciling definitions, checking for cycles and inconsistencies, choosing preferred terms, disambiguating homonyms, among others.

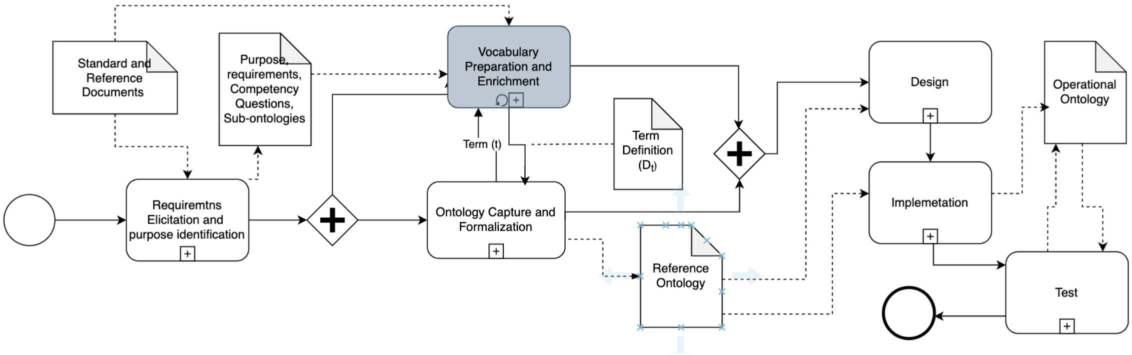

Standard and Reference Documents
Purpose, requirements, Competency Questions, Sub-ontologies
Vocabulary Preparation and Enrichment
Term (t)
Term Definition (D_t)
Requiremtns Elicitation and purpose identification
Ontology Capture and Formalization
Reference Ontology
Design
Implemetation
Operational Ontology
Test

Figure 1: Ontology Engineering Process

Moreover, especially in the case of cybersecurity domain, there are many reference and terminological documents to take into account, which turn the Vocabulary construction into an even more expensive and time-consuming task. In the present work, we intend to address some of these problems in an agile way, by proposing a Vocabulary preparation and enrichment sub-process using a RAG/LLM approach.

2.2. RAG

Based on insights derived from references [7, 8, 9, 10, 11], Retrieval-Augmented Generation (RAG) is a technique that enhances text generation in large language models (LLMs) by integrating information from external, private, or proprietary data sources that are reliable, up-to-date, and provide additional context. This approach improves the accuracy and reliability of the model's output by referencing an external knowledge base before generating a response. Additionally, RAG promotes transparency by linking the generated text to specific, relevant sources, offering users insight into the model's generative process.

The RAG technology has been developing rapidly. RAG originated alongside the Transformers framework, designed to enhance text generation by incorporating external context [12]. Subsequently, the technology focused on prioritizing the most relevant information to enhance LLM responses. RAG was later utilized to assist in fine-tuning LLMs, further improving their contextual awareness and accuracy [13]. Thus, there are different types of RAG, which can be classified based on their implementation, architecture, or approach to integrating external data. Among the various types of approaches, the following stand out [14, 15, 16]:

- **Naïve RAG**: Documents are segmented into smaller chunks and transformed into vector embeddings for efficient storage. When a query is received, the system retrieves the most relevant chunks using a similarity function and provides them as context for a language model.
- **Advanced RAG**: enhances the base model with sophisticated preprocessing (e.g., query reformulation) and post-processing (e.g., document re-ranking) to optimize retrieval accuracy. It can integrate LLMs, Large Retrieval Models (LRMs), and Small Language Models (SLMs) to generate precise, coherent, and contextually enriched responses.
- **Modular RAG**: Allows individual components to be replaced or fine-tuned independently, including the retriever (which fetches relevant data), the processor (which pre-processes information), and the generator (which leverages an LLM or LRM to produce coherent and contextually accurate text).
- **Corrective RAG**: After generating a response, the system cross-references it with trusted data sources to identify and correct inaccuracies, ensuring greater factual accuracy and reliability.
- **Speculative RAG**: Typically utilizes an LLM or LRM to generate plausible responses by combining retrieved information with pattern-based reasoning and informed assumptions.

2.3. Information Retrieval and Ranking Functions

Information Retrieval (IR) [17, 18] is an area concerned with the extraction of desired information from various sources, but mainly textual content. The most used models in IR are vector space models and probabilistic models, although there are techniques that do not involve embedding of text, such as ranking functions.

Vector space models[19, 20] aim to *embed* words (or even entire fragments of text) as vectors in a high-dimensional space. *Relevance* or *similarity* are then translated to *distance* between such vectors, which can be represented in many ways, but mostly via the dot-product between two vectors.

Probabilistic models are based on the principle that documents in a corpus should be ranked by decreasing probability of their relevance to a queried term - the Probabilistic Ranking Principle [21]. Different implementations provide their own probability estimation techniques, and each domain requires adequate crafting of them.

Ranking Functions are another way to retrieve information from texts - usually paired together with word embeddings. Given a set (usually called a *corpus*) of fragments of text (*documents*)[1] and a certain term to be queried in these documents, this function creates a score for each - indicating the relevance of each to the queried term.

Tf-idf[22] is a ranking function that stands for a mix of *term frequency* and *inverse document frequency*. It is, in fact, a way to take both these concepts into account when scoring documents regarding certain queried terms. Term frequency is the amount of times the queried term appears in each document. Inverse document frequency is the logarithm of the inverse of the frequency with which the term appears in all of the documents - indicating how relevant the document in question is compared to the others. These scores are, then, multiplied (let N be the number of documents, d a document and q a queried term):

$$\text{TF-IDF}(d, q) = \text{tf}(d, q) \cdot log\left(\frac{N}{\text{df}(d)}\right)$$

Over decades, Tf-idf has been used in several applications (e.g.[23]), but it does have certain limitations. It does not consider concepts such as normalization and saturation. Normalization is related to the size of each document, i.e. if a term appears 10 times in a document that has 100 words, it should be more relevant than appearing 11 times in 1000 words. Saturation indicates that there must be a threshold up to which the term is still relevant in the document - the more it appears above this point, the less important the syntactical presence of the term in the document is to its semantics.

BM25[24] stands for *Best Match 25* (25 indicates that it was the 25th iteration of the refinement of this algorithm). It is a series of improvements over Tf-idf considering concepts such as the ones described above:

$$\text{BM25}(d, q) = \sum_{t \in q} log\left(\frac{N}{\text{df}(t)}\right) \cdot \frac{(k_1 + 1) \cdot \text{tf}(t, d)}{k_1 \cdot \left((1 - b) + b \cdot \frac{len(d)}{len_{avg}}\right) + \text{tf}(t, d)}$$

$$k_1, b \text{ - parameters}$$
$$len(d) \text{ - size of document } d$$
$$len_{avg} \text{ - average document size}$$

There have been many implementations of BM25 over the years [25, 26, 27], each adjusting parameters, considering different domains, and considering different linguistic concepts. However, most of them have better performance than plain Tf-idf in most scenarios.

[1] In this section we will utilize these terms to define ranking functions, but they are confusing when taking our implementation into account. In the following sections, we will refer to a *corpus* as a *document*, and the documents in it exclusively as *fragments of text* or *sentences*.

2.4. Cybersecurity

Cybersecurity, the domain addressed in this article, is the prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation [28]. The ontological analysis of the cybersecurity domain is highly valued, as it reveals how different initiatives model reality in distinct ways, which can directly impact the understanding of the domain, the interoperability of security systems, and the subsequent implementation of policies and actions. We will extract fundamental concepts from these initiatives; therefore, we provide a brief description of each.

Several initiatives have been proposed to guide best practices in cybersecurity, providing guidelines for identifying, preventing, and responding to threats. Some of the most prominent cybersecurity frameworks and standards include the MITRE strategies, NIST approaches, ISO/IEC 27001, CIS Controls, COBIT, and OWASP.

MITRE is an American corporation that has developed two complementary frameworks: MITRE ATT&CK and MITRE D3FEND. The philosophy of MITRE ATT&CK is to move from a reactive defense to a proactive defense based on adversary behavior [29]. Therefore, they focus on how adversaries achieve their objectives, i.e., the tactics, techniques, and procedures (TTPs) employed in real-world attacks. MITRE D3FEND, on the other hand, focuses on defensive countermeasures mapped onto ATT&CK offensive tactics and techniques [30]. Once one knows how adversaries attack, one can detail how to defend against them.

NIST (National Institute of Standards and Technology) produces a wide range of publications, standards, and frameworks related to cybersecurity. The NIST Cybersecurity Framework (CSF) is a high-level strategic guide for organizations to manage and mitigate cyber risk [31]. It is organized into six main functions: Govern (establishing strategy, expectations, and policies), Identify (mapping assets, risks, and vulnerabilities), Protect (implementing controls such as encryption and access management), Detect (continuous monitoring to identify incidents), Respond (executing mitigation and communication plans), and Recover (restoring services and improving resilience). The CSF acts as an aggregator, pointing to other standards and norms for the implementation details. The NIST Special Publication (SP) 800 Collection is the most comprehensive collection of NIST cybersecurity guidelines and recommendations, such as NIST 800-53, a catalog of security controls that organizations should implement [28]. This extensive catalog provides technology-agnostic controls used by risk teams, while security operations teams more commonly use MITRE mitigations. It is possible to map between the two to find a common language. The glossary of terms [32] used in all technical publications is useful for ontology design.

ISO/IEC 27001 [33], developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), is an international standard that defines requirements for an Information Security Management System (ISMS), focusing on the confidentiality, integrity, and availability of data. It covers risk management and security controls organized into domains (such as policies, asset management, incident response, and certification). An organization may seek an ISO 27001 certification to demonstrate compliance with an international ISMS standard, which often covers a large portion of the NIST CSF requirements and implements many of the controls in NIST SP 800-53.

CIS Controls [34], developed by the Center for Internet Security (CIS), provide a set of practical and prioritized controls to protect systems and data. It is focused on immediate implementation, making it accessible for organizations seeking quick and effective solutions. The CIS Controls are closely aligned with the NIST approaches, sharing common goals and complementary structures.

Control Objectives for Information and Related Technology (COBIT), developed by the Information Systems Audit and Control Association [35], is an IT governance framework that combines cybersecurity with organizational strategic objectives. It promotes effective technology and information management, aligning security with business processes.

OWASP (Open Web Application Security Project) [36] is an organization focused on software security, particularly in web applications and APIs, providing guidelines and tools for developers and businesses. Its most well-known project, the OWASP Top 10, lists the primary vulnerabilities in web applications, such as SQL injection and authentication failures, helping prioritize mitigation actions.

These initiatives, each with its specificities, are complementary and can be combined to create robust cybersecurity strategies tailored to the needs and objectives of each organization. In the context of ontologies, they provide systematic frameworks that organize, standardize, and implement security concepts consistently and interoperable, facilitating the development and maintenance of ontologies.

3. Related Work

A literature mapping was conducted by analyzing 19 selected articles, following the principles and phases proposed by Kitchenham (2004) [37]. We explored well-known scientific databases, including Scopus, Web of Science, IEEE Xplore, SpringerLink, and Google Scholar. Our review identified key challenges and opportunities for using AI approaches, such as large language models (LLM) and retrieval-augmented generation (RAG) to support ontology development. We analyze works addressing various aspects of the ontology development process, including studies focused specifically on vocabulary concept harmonization. Of the related works reviewed, 12 studies focused on building or improving ontologies using LLM during the initial phases of ontology engineering, such as specification and conceptualization.

LLMs are proven to be a promising approach for ontology learning and engineering, as they combine efficient extraction of structured knowledge from natural text with human collaboration for refinement and validation. Techniques for automatic discovery of taxonomic relationships and dynamic generation of ontological components using RAG have been proposed in recent studies. Giglou et al. (2023) [4] propose an approach that uses LLMs for Ontology Learning (OL). The authors investigate whether LLMs can automatically extract and structure knowledge from natural language text. Nine LLM families were evaluated for three main OL tasks: term typing, taxonomy discovery, and extraction of non-taxonomic relations. Doumanas et al. (2024) [38] present an LLM-enhanced ontology engineering (OE) approach, aiming to identify how OE tasks can be completed with LLM and human collaboration. LLMs are employed to generate domain ontologies for modeling Search and Rescue (SAR) missions in wildfire incidents. The authors analyze LLM capabilities to OE and evaluate the human-machine synergy to represent knowledge, focusing on the SAR domain. Toro et al. (2024) [39] present an ontology generation method employing LLM and RAG (DRAGON-AI) aiming at generating textual and logical ontology components. According to the authors, DRAGON-AI has high precision for relationship generation, but has slightly lower precision than from logic-based reasoning; evaluators with the highest level of confidence in a domain were better able to discern flaws in AI-generated definitions. Mateiu and Groza (2023) [40] enrich ontologies by translating Natural Language (NL) into Description Logic. A GPT model is fine-tuned to convert NL into OWL. Pairs of sentences in NL and the corresponding translations for fine-tuning are designed. The training pairs cover aspects of ontology engineering, such as instances, domain and range of relations, and object property relationships. The resulting axioms were used to enrich an ontology, supervised by human experts.

Abolhasani and Bran (2025) [41] propose the OntoKGen platform, which uses LLMs to extract ontologies from technical texts and create new branches of these ontologies through user interaction and validation to define concepts, relationships, and properties. Based on the confirmed ontology, the platform generates KG in an automated, interactive, and adaptive manner, reducing user intervention while allowing necessary adjustments. In future work, the authors propose to integrate the OntoKGen-generated KGs into RAG systems, enabling dynamic data manipulation via an interface. Although OntoKGen does not directly work with RAG systems, it represents an advance in the extraction and creation of ontologies using LLMs.

Vrolijk et al. (2023) [42] present an ontology learning system for the job market based on the ESCO ontology, which uses LLM combined with RAG techniques to extract, classify and relate mentions of skills and occupations from online job advertisements. The system proposes a three-layer architecture that integrates automatic processing and human interaction to keep the ontology updated, identifying new entities and relationships. The experiments indicate that the method improves performance in extracting mentions, classifying relationships, discovering knowledge, and suggesting new entities for ontology extension.

Bran et al. (2025) [43] introduce the OntoRAG methodology, which combines LLM with ontologies to enhance knowledge generation in scientific domains, where the goal is to mitigate "hallucination" problems. The approach was tested on a benchmark in the Single Atom Catalysis (SAC) domain, showing its effectiveness in predicting synthesis procedures. The results indicate that OntoRAG outperforms traditional RAG methods, highlighting the potential of integrating ontologies as knowledge representation alongside LLM models. In addition, the authors present the OntoGen tool, which allows the automatic generation of ontologies from documents. The OntoGen process can be divided into three stages, namely: (a) vocabulary extraction; (b) category generation; and (c) taxonomy extraction. These stages facilitate the adaptation of the OntoRAG method when applied to new domains. Despite the results presented, the authors note that user supervision is still needed when creating ontologies.

LLMs have supported the expansion and enrichment of ontologies, demonstrating effectiveness in the automated generation of ontological components (e.g., competency questions and RDF mappings) and the structured extraction of knowledge from unstructured texts, with applications in diverse domains. Yang et al. (2024) [44] propose an LLM-based ontology expansion method. LLMs are used to formulate competency questions (CQs) and to extend the initial ontology. The authors created a knowledge graph for breast cancer treatment. Mukanova et al. (2024) [45] propose an LLM-powered NLP method for ontology enrichment. The authors aim to process natural language texts and extract data from the text that matches the semantics of an ontological model. LLM extracts data from a Web page and converts it into lists with information relevant to an ontology. The proposed method is implemented using the example of an ontological model that describes a geographical configuration. Val-Calvo et al. (2025) [46] use LLM to aid in the development of ontology from data sets, increasing automation of ontology-based KG generation. The authors developed an LLM method to enhance ontology engineering through data pre-processing, ontology planning, building, and entity improvement. The proposed method can generate mappings and RDF data, but the authors focus on ontologies.

LLMs in conjunction with semi-automated approaches can support KG and ontology engineering, e.g., formulating CQs, developing or evaluating KGs with lower human intervention, and enabling conversational frameworks for eliciting requirements in ontologies. Kommineni et al. (2024) [47] present an LLM-supported approach for semi-automatically building an ontology and KG. The proposed approach involves: i) formulating competency questions (CQs); ii) developing an ontology (TBox) based on these CQs; iii) constructing KGs using the developed ontology; and iv) evaluating the resultant KG with minimal to no involvement of human experts. To evaluate the answers generated via RAG and the KG concepts automatically extracted using LLMs, the authors designed a judge LLM that rates the generated content. Zhang et al. (2024) [48] present a framework for conversational ontology engineering (OntoChat), aimed at supporting requirement elicitation, analysis, and testing. OntoChat aids users in creating user stories and extracting competency questions. The authors replicated the engineering of the Music Meta Ontology and collected preliminary metrics on the effectiveness of each component from users.

Our approach (LAVOHA) is focused on defining concepts from cybersecurity vocabularies and ontologies. We harness LLM to analyze and harmonize concept definitions in natural language and propose term relationships of a security incident response glossary. LAVOHA supports ontologists in the specification and conceptualization phases of the ontology engineering process. LAVOHA shares similarities with other works, particularly in leveraging LLMs for ontology-related tasks. Unlike related works, our method employs LLMs to process natural language and extract structured knowledge, supporting ontology engineering. The focus on automating parts of the ontology development process aligns with [46], [38], and [47], which also aim to reduce human effort through LLM assistance. Similarly to [44] and [47], LAVOHA involves defining and refining concepts (in our case, cybersecurity terms) using LLM-generated insights. The emphasis on human-AI collaboration is another shared aspect, as seen in [38] and [48], where human expertise guides and validates the LLM outputs.

Although related work (e.g. [4], [45], [39], [40]) focuses on general ontology learning or enrichment, our approach is domain-specific, targeting cybersecurity vocabularies and incident response terminology.

Unlike [39] and [40], which use LLMs for the generation or translation of logical axioms, our method focuses on conceptual harmonization and proposal of relationships, supporting the early stages of the ontology engineering. [47] and [48] present automated evaluation mechanisms (e.g., judge LLM or conversational interfaces), whereas our approach prioritizes ontologist-guided refinement rather than full automation. This distinguishes our work from [47], which minimizes human involvement, and aligns more closely with the emphasis on human-machine synergy from [38]. Finally, [44] and [46] integrate ontologies with KGs, and our current scope is limited to glossary and ontology conceptualization, although future extensions could explore KG integration. Our approach shares foundational LLM-based strategies with other works, but distinguishes itself through its cybersecurity focus, conceptual harmonization goals, and balanced human-AI collaboration. Despite the large number of studies on the use of LLMs in supporting ontology development, the specific issue of concept harmonization has been little explored in the literature. Table 1 presents a comparative analysis of related work.

Table 1
Related Work

Ref.	**Purpose**	**Input**	**Output**	**Application Domain**
[4]	Extracting and structuring knowledge	NL text	Structured knowledge	Geographic and medical
[44]	Formulating CQ and extending ontology	Ontology	CQ and extended ontology	Breast cancer treatment
[45]	Processing natural language texts	NL text and ontological model	Information lists relevant to a domain ontology	Geographic
[46]	Automating ontology-based KG generation	Datasets and ontology	Ontology-based KG	Commercial activities
[38]	Identifying how OE tasks can be completed with LLM	NL text	Domain ontology	SAR missions in wildfire incidents
[39]	Generating textual and logical ontology components	NL text	Ontology components	Basic Science (e.g., Gene, Biological, Environment)
[40]	Translating NL into DL	NL text	OWL ontology	Family relations (e.g., father, sister, etc.)
[41]	Extraction and creation of ontologies and generation of knowledge graphs	Complex technical documents (text)	Ontology and KG	Reliability and Maintenance in semiconductor manufacturing equipment
[42]	Automatic learning and updating of ontologies	Job advertisements (text)	Ontology components	Labor market, job ontologies
[43]	Reduction of hallucinations with ontologies and automatic generation of ontologies	Responses generated with greater accuracy and relevance and ontology components	Accurate and relevant responses and ontology components	Single Atom Catalysts (SAC)
[47]	Semi-automatically building ontology and KG	CQs	Ontology and KG	DL methodologies
[48]	Supporting requirement elicitation, analysis, and testing	User stories	CQs	Music metadata
This Work	*Harnessing LLM to analyze and harmonize concept definitions in NL text*	*NL text*	*Glossary of a domain ontology*	*Cybersecurity Incident Response*

4. The LAVOHA Method

This section introduces *LAVOHA*, a simplified version of an Advanced RAG-inspired method (see subsection 2.2) designed to support vocabulary harmonization in the ontology creation process.

4.1. Conceptual Description

Given a term t to be incorporated into a vocabulary, together with $Q_t = \{q_1, q_2, \ldots, q_{|Q_t|}\}$[2], a set of *queried terms* (i.e. words related to the definition of term t), and a corpus $D = \{d_1, d_2, \ldots, d_{|D|}\}$ consisting of relevant documents within the target domain, *LAVOHA* extracts relevant sentences from

[2] In this article, $|X|$ denotes the cardinality of any arbitrary set X

documents in D and uses them to query an LLM for suggested definitions for t. It is important to highlight that if t is a single word term then $t \in Q_t$, so that the algorithm will return the sentences related to t itself. In case t is a compound name, then each word in t must be in Q_t. Figure 2 presents a modular overview of the designed method.

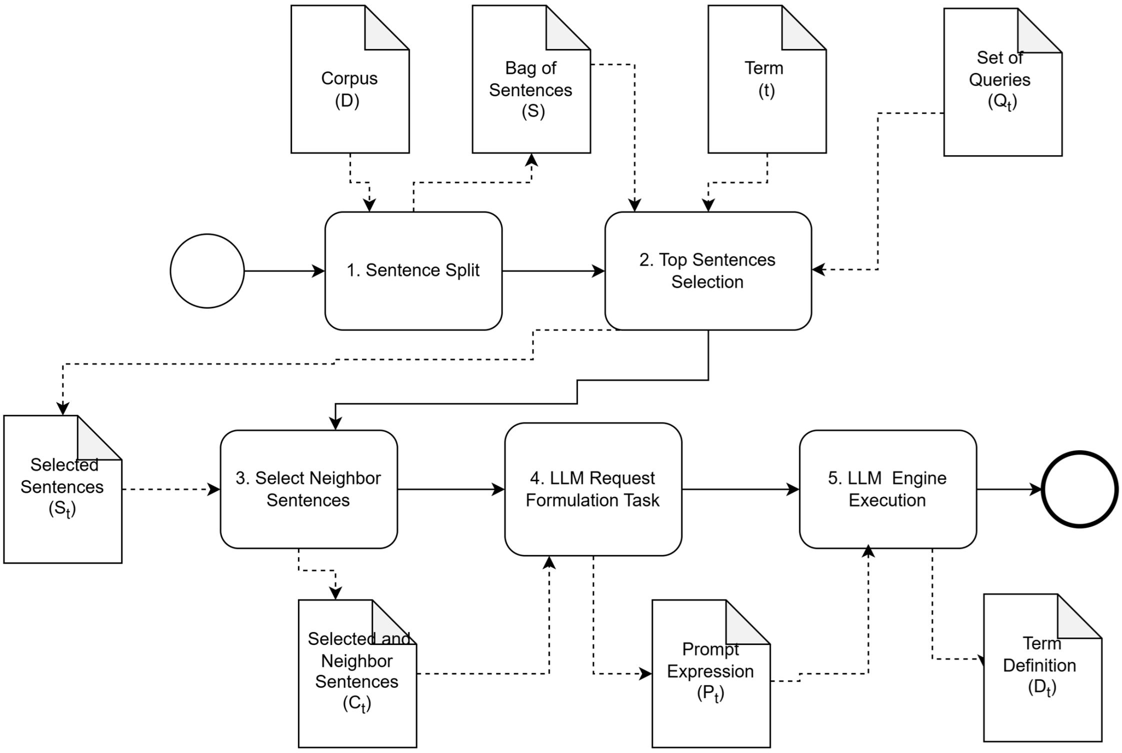


Figure 2: Vocabulary Harmonization Process - Modular Overview

Step 1 (Sentence Split) splits each document $d \in D$ into sentences, yielding a *Bag of Sentences* S, the set of all sentences in D. In *Step 2 (Top Sentences Selection)*, given S, t and Q_t, this step evaluates a relevance score for each pair $(q, s) \in Q_t \times S$. For each $q \in Q_t$, the top N sentences are returned, resulting in a set S_t of up to $|Q| \times N$ distinct sentences. Then, for each sentence $s_t \in S_t$, *Step 3 (Select Neighbor Sentences)* retrieves the M sentences that precede s_t and the M sentences that succeed s_t in d_{s_t}. The output of this step is, therefore, a set C_t of up to $|S_t| \times (2M + 1)$ distinct sentences, since the original sentence in S_t is also included. The next step *(Step 4 - LLM Request Formulation Task)* constructs P_t, the prompt expression for an LLM L chosen by the user. To this end, the following strings are concatenated: *"You are a specialist in "; name-of-the-domain-of-the-application; ". "; "Define the term "; t; ", based on the definitions stated in the following set: "; C_t*. Note that *name-of-the-domain-of-the-application*, t, and S_t are variables and are not strings themselves. Indeed, they contain the strings to be concatenated to form the prompt. Finally, *LAVOHA's* last step *(Step 5 - LLM Engine Execution)* consists of the execution of $L's$ engine, given the prompt P_t. Its output is D_t, the definition for t suggested by L.

4.2. Implementation Details

We provide further details on the implementation of the most challenging steps of the LAVOHA method. The LAVOHA implementation artifacts are available in the GitHub repository [3]. The system was developed in Python (version 3.12), due to its vast availability of useful packages and documentation online.

[3] https://github.com/anonymous_for_double_blind_revision

Pre-processing

Before Sentence Split, the pdf documents are converted to text files. This process is done only once per document, as a pre-processing step, since it is highly time-costly. Whenever a new document is added to the pool, it is quickly converted to a text file. To convert `pdf` files to `txt`, we make use of the `PyMuPdf`[4] in version 1.25.5.

Step 1 (Sentence Split)

In order to properly break the textual content of each file, we make use of the `nltk`[5] NLP package version 3.9.1, providing the appropriate language identifier depending on the document (`"english"` for documents in English, and `"portuguese"`, for Portuguese) in order to find and properly ignore the correct stopwords. The results presented in section 5 concern only the english version of the documents.

Step 2 (Top Sentences Selection)

The selection of the best sentences was made using the BM25 algorithm. Our choice of BM25 implementation is the package `rank-bm25`[6], in version 0.2.2.

The definitions of the BM25 parameters for each term are in a module (`queries.py`), containing a dictionary with information pertaining to the BM25 parameters for the 37 terms initially assigned to the glossary. This module centralizes and makes it easier to change and adapt BM25 parameters should the result of a BM25 call be unsatisfactory - this happens mostly in situations where not enough relevant sentences are selected, or when many irrelevant sentences are selected via the algorithm. The number of sentences retrieved in this module can be set in a property file.

Step 3 (Neighbor Sentences Selection)

The number of neighbors recovered in this step can be adjusted via a property file. In the experiment reported in the present paper, the number of neighbors was set to 0 (no neighbors), since only in special cases does the context surrounding the selected sentences provide useful information. This evaluation should be conducted individually for each case.

5. Case Study

We are developing an ontology for the cybersecurity domain, following a 4-phase methodology: specification, conceptualization, formalization, and implementation. In this methodology, the glossary of terms should be drafted preliminarily during specification and finalized during conceptualization. During the conceptualization phase, we faced the difficulty of reconciling different definitions of terms originating from documents assigned as knowledge sources. At this point, we created the LAVOHA method to support this task. This section presents the results of the LAVOHA method applied to this difficulty.

5.1. Experiment Configuration

As the document corpus D for the cybersecurity domain, we used the following set of documents:

- ATTACK_Design_and_Philosophy_March_2020.pdf
- getting-started-with-attack-october-2019.pdf
- mitre_TTPs.pdf
- NBRISO-IEC 27035.pdf

[4] https://pypi.org/project/PyMuPDF/
[5] https://www.nltk.org/
[6] https://pypi.org/project/rank-bm25/

- NIST.SP.800-61r2.pdf

The following list shows each term t used in the experiment, together with its set of related words (queries) Q_t. It is worth recalling that the terms and their respective related words, defined by the user, are precisely the concepts that require harmonization, since the meanings may vary in each document.

- Attack: "attack", "attempt", "access", "damage", "interrupt", "malicious", "degrade", "destroy",
- Attack vector: "vector", "attack", "method", "methods", "technique", "techniques", "access"
- Campaign: "campaign", "grouping", "activities", "intrusion", "period", "targets", "objectives"
- Damage: "damage", "effect", "event", "incident", "occurrence", "loss"
- Event: : "event", "occurrence", "observable", "indication", "incident", "suspicion", "adverse"
- Incident: "incident", "occurrence", "risk", "confidentiality", "integrity", "availability", "information", "violation", "threat", "policies"
- Information Asset: "asset", "information", "value", "person", "organization", "organisation", "medium", "resource", "critical"

The configuration properties used in the experiment were the following:

- N = 6. Step 2 selects the top six sentences.
- M = 0. Step 3 adds no neighbors.
- $L \in \{GPT\text{-}4o, DeepSeek\text{-}V3\}$. In Step 6 we used both GPT-4o and DeepSeek-V3.

5.2. Results

To assess whether LAVOHA indeed improved the LLM output, we generated the output (reconciled definitions) in three different ways:

- Using only the LLM, without LAVOHA;
- Using the LLM with LAVOHA;
- Through human discussions until reaching a consensual definition.

The human consensual definition was considered the appropriate response, and it was later compared to the first two definitions to evaluate whether the use of LAVOHA improved the LLM response for this task. To achieve this, we calculated the sentence embedding as the average of its word vectors, the cosine similarity between the responses generated by LLMs, and the consensus definition. The definition embeddings were computed with the pre-trained FastText library [49, 50].

Table 2 shows the mean value and standard deviation of the cosine similarity over the seven terms' definition. We note that with the use of LAVOHA, the quality of GPT4-o output improved from a 0.9261 mean similarity with the consensus definition to 0.9317, whereas DeepSeek improved from 0.9279 to 0.9443. Therefore, the DeepSeek model responded better with the help of LAVOHA. This result may be due to DeepSeek having a smaller training set than GPT4-o.

Table 2
LLM's definitions mean similarity to consensual definitions

	Mean	**SD**
GPT-4o	0.9261	0.0516
DeepSeek-V3	0.9279	0.0242
GPT-4o+LAVOHA	0.9317	0.0519
DeepSeek-V3+LAVOHA	0.9443	0.0364

We performed the automated evaluation because it is difficult to analyze manually the nuances that differentiate the two definitions. Still, it may be useful to present definitions produced with and without LAVOHA to illustrate the quality improvement achieved by the method. As an example, we list two definitions of attack: the first was generated by DeepSeek-V3 alone, and the second by DeepSeek-V3 assisted by LAVOHA:

1. A deliberate, malicious attempt by an individual, group, or system to exploit vulnerabilities in a computer system, network, or digital infrastructure with the intent to: - Compromise confidentiality, integrity, or availability (CIA triad) of data or services. - Gain unauthorized access, disrupt operations, steal information, or cause harm. - Deploy malware, execute code, manipulate systems, or escalate privileges. Attacks can be active (directly altering or damaging systems) or passive (eavesdropping without modification). Common types include phishing, ransomware, DDoS, SQL injection, and zero-day exploits.
2. A deliberate, malicious attempt by an adversary to compromise, disrupt, or destroy the confidentiality, integrity, or availability of systems, networks, or data. Attacks may employ various tactics, techniques, and procedures (TTPs), such as exploiting vulnerabilities, deploying malware, or leveraging social engineering, to achieve objectives like unauthorized access, data theft, service disruption (e.g., Denial of Service), or system destruction. These actions can target technical infrastructure (e.g., endpoints, cloud resources) or human elements (e.g., phishing), and often mimic normal activity to evade detection.

The second definition feels cleaner and more precise. In the first sentence, it defines attack, in the second, it demonstrates how attacks may be performed, and in the third sentence, it enumerates the possible targets of an attack. Furthermore, it uses the verb "target", an important predicate, because it characterizes the relation between a typical attack and the attacked assets, suggesting a possible triple modeling (Attack, targets, Asset). Then it identifies "technical infrastructure" as a target, a general term that encompasses the vulnerable entities listed in the first definition. It also includes "human elements", which is a conceptual gap in the first definition.

The other six term definitions generated by DeepSeek and the seven term definitions generated by GPT presented similar improvement with the aid of LAVOHA. Furthermore, DeepSeek answers for the seven terms without LAVOHA seem to vary little, as if reading from the same source. For instance, DeepSeek mentions "confidentiality, integrity, or availability (CIA triad)" in 4 of the 7 definitions without LAVOHA. However, when assisted by LAVOHA, it mentions the CIA triad in only one of the definitions, which shows a better separation of concepts by the different definitions.

6. Conclusion

We introduced LAVOHA, a method designed to harmonize conflicting concept definitions, specifically applied within cybersecurity vocabularies and ontologies. The approach leverages natural language analysis to produce unified consensus-based definitions of security concepts, as shown through a case study that compares LAVOHA-generated definitions with human consensus. The results favor LAVOHA over relying solely on large language models (LLMs), showing improved suitability.

As future work, we intend to extend our approach to other phases of Ontology Engineering, investigating how LLMs can assist in writing an ontology and whether LAVOHA-like methods can enhance their performance. Regarding the focus of the present work, namely vocabulary harmonization, future research could explore additional quantitative evaluation methods, such as measuring the perplexity of the consensus definition when submitted to the LLM in different scenarios. We expect that better-equipped LLMs (possibly enhanced by LAVOHA) will exhibit lower perplexity scores for the consensual definition. It would also be worthwhile to test alternatives to the BM25 algorithm for sentence selection, such as using LLM-based embeddings to represent queries and candidate sentences in Step 2 of LAVOHA.

Acknowledgments

This work was supported by the Brazilian Funding Authority for Studies and Projects (FINEP) trough the projects [CyberSemantics - Contract No. 0.1.22.0335.00/Ref. FINEP No. 0172/22] and [S2C2 - Contract No. 0.1.20.0272.00/Ref. FINEP No. 2904/20] and the Systems Development Center of the Brazilian Army. E.H.Haeusler was partially suported by CNPq grant 309287/2023-5.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o for grammar and spelling checks and text translations. After using this tool, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

[1] R. de Almeida Falbo, Sabio: Systematic approach for building ontologies, in: G. Guizzardi, O. Pastor, Y. Wand, S. de Cesare, F. Gailly, M. Lycett, C. Partridge (Eds.), Proceedings of the 1st Joint Workshop ONTO.COM / ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems, ONTO.COM/ODISE@FOIS 2014, Rio de Janeiro, Brazil, September 21, 2014, volume 1301 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014. URL: https://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf.

[2] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology framework: A scenario-based methodology for ontology development, Applied Ontology 10 (2015) 107–145. URL: https://journals.sagepub.com/doi/abs/10.3233/AO-150145. doi:10.3233/AO-150145.

[3] N. J. Mariano Fernández-López, Asunción Gómez-Pérez, Methontology: From ontological art towards ontological engineering, Miscellaneous (1997).

[4] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: International Semantic Web Conference, Springer, 2023, pp. 408–427.

[5] A. De Nicola, M. Missikoff, A lightweight methodology for rapid ontology engineering, Commun. ACM 59 (2016) 79–86. URL: https://doi.org/10.1145/2818359. doi:10.1145/2818359.

[6] P. M. L. Scheidegger, M. L. M. Campos, M. C. Cavalcanti, An approach for systematic definitions construction based on ontological analysis, in: E. Garoufallou, S. Virkus, R. Siatri, D. Koutsomiha (Eds.), Metadata and Semantic Research - 11th International Conference, MTSR 2017 Tallinn, Estonia, November 28 - December 1, 2017, Proceedings, volume 755 of *Communications in Computer and Information Science*, Springer, 2017, pp. 87–99. URL: https://doi.org/10.1007/978-3-319-70863-8_9. doi:10.1007/978-3-319-70863-8_9.

[7] Amazon Web Services, What is RAG? - Retrieval-Augmented Generation explained, https://aws.amazon.com/what-is/retrieval-augmented-generation/, 2023. Acessado em 10 de outubro de 2025.

[8] IBM Research, What is retrieval-augmented generation (RAG)?, https://research.ibm.com/blog/retrieval-augmented-generation-RAG, 2023. Acessado em 10 de outubro de 2025.

[9] Intel, What is RAG? Retrieval-Augmented Generation explained, https://www.intel.com/content/www/us/en/learn/what-is-rag.html, 2025. Acessado em 10 de outubro de 2025.

[10] Elastic, What is retrieval-augmented generation?, https://www.elastic.co/what-is/retrieval-augmented-generation, 2025. Acessado em 10 de outubro de 2025.

[11] Oracle, What is Retrieval-Augmented Generation (RAG)?, https://www.oracle.com/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/, 2023. Acessado em 10 de outubro de 2025.

[12] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[13] Y. Gao, et al., Retrieval-augmented generation for large language models: A survey, https://arxiv.org/abs/2312.10997, 2024. Acessado em 10 de outubro de 2025.

[14] IBM, What are RAG techniques?, https://www.ibm.com/think/topics/rag-techniques, 2025. Acessado em 10 de outubro de 2025.

[15] Weka, What is Retrieval-Augmented Generation (RAG)?, https://www.weka.io/learn/guide/ai-ml/retrieval-augmented-generation/, 2024. Acessado em 10 de outubro de 2025.

[16] S. Homayoun, 6 types of Retrieval-Augmented Generation (RAG) techniques you should know, https://homayounsrp.medium.com/6-types-of-retrieval-augmented-generation-rag-techniques-you-should-know-b45de9071c79, 2024. Acessado em 10 de outubro de 2025.

[17] A. Singhal, Modern information retrieval: A brief overview, IEEE Data Eng. Bull. 24 (2001) 35–43.

[18] K. Hambarde, H. Proença, Information retrieval: Recent advances and beyond, IEEE Access 11 (2023) 76581–76604. doi:10.1109/ACCESS.2023.3295776.

[19] A. Neelima, S. Mehrotra, A comprehensive review on word embedding techniques, in: 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), 2023, pp. 538–543. doi:10.1109/ICISCoIS56541.2023.10100347.

[20] L. Gutiérrez, B. Keith Norambuena, A Systematic Literature Review on Word Embeddings: Proceedings of the 7th International Conference on Software Process Improvement (CIMPS 2018), 2019, pp. 132–141. doi:10.1007/978-3-030-01171-0_12.

[21] A. de Vries, A. Wilschut, On the integration of ir and databases, in: Database issues in multimedia, 1999, pp. 16–31.

[22] G. Salton, C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, Technical Report, USA, 1987.

[23] R. A. Shahzad Qaiser, Text mining: Use of tf-idf to examine the relevance of words to documents, International Journal of Computer Applications 181 (2018) 25–29. URL: https://ijcaonline.org/archives/volume181/number1/29681-2018917395/. doi:10.5120/ijca2018917395.

[24] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at TREC-3, in: D. K. Harman (Ed.), Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994, volume 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1994, pp. 109–126. URL: http://trec.nist.gov/pubs/trec3/papers/city.ps.gz.

[25] X. H. Lù, Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024. URL: https://arxiv.org/abs/2407.03618. arXiv:2407.03618.

[26] A. Trotman, A. Puurula, B. Burgess, Improvements to bm25 and language models examined, in: Proceedings of the 19th Australasian Document Computing Symposium, ADCS '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 58–65. URL: https://doi.org/10.1145/2682862.2682863. doi:10.1145/2682862.2682863.

[27] M. Taylor, H. Zaragoza, N. Craswell, S. Robertson, C. Burges, Optimisation methods for ranking functions with multiple parameters, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM '06, Association for Computing Machinery, New York, NY, USA, 2006, p. 585–593. URL: https://doi.org/10.1145/1183614.1183698. doi:10.1145/1183614.1183698.

[28] National Institute of Standards and Technology, Security and Privacy Controls for Information Systems and Organizations: NIST Special Publication 800-53, Revision 5, Technical Report, National Institute of Standards and Technology, Gaithersburg, 2020.

[29] B. E. Strom, A. Applebaum, D. P. Miller, K. C. Nickels, A. G. Pennington, C. B. Thomas, MITRE ATT&CK: Design and Philosophy, Technical Report, MITRE Corporation, 2020. URL: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf.

[30] P. E. Kaloroumakis, M. J. Smith, Toward a Knowledge Graph of Cybersecurity Countermeasures, Technical Report, MITRE Corporation, 2023. URL: https://d3fend.mitre.org/resources/D3FEND.pdf.

[31] National Institute of Standards and Technology, Cybersecurity framework, 2025. Available at: https://www.nist.gov/cyberframework, accessed on May 28, 2025.

[32] Computer Security Resource Center, Glossary, 2025. URL: https://csrc.nist.gov/glossary.

[33] International Organization for Standardization, Iso/iec 27001:2022 – information security, cybersecurity and privacy protection, 2022. Available at: https://www.iso.org/standard/27001, accessed on May 28, 2025.

[34] Center for Internet Security, Cis critical security controls, 2025. Available at: https://www.cisecurity.org/controls, accessed on May 29, 2025.

[35] ISACA, Cobit 2019 framework: Introduction and methodology, 2019. Available at: https://www.isaca.org/resources/cobit, accessed on May 29, 2025.

[36] OWASP, Owasp top ten, 2025. Available at: https://owasp.org/www-project-top-ten/, accessed on May 29, 2025.

[37] B. Kitchenham, Procedures for performing systematic reviews, Keele, UK, Keele University 33 (2004) 1–26.

[38] D. Doumanas, A. Soularidis, K. Kotis, G. Vouros, Integrating llms in the engineering of a sar ontology, in: IFIP International Conference on Artificial Intelligence Applications and Innovations, Springer, 2024, pp. 360–374.

[39] S. Toro, A. V. Anagnostopoulos, S. M. Bello, K. Blumberg, R. Cameron, L. Carmody, A. D. Diehl, D. M. Dooley, W. D. Duncan, P. Fey, et al., Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai), Journal of Biomedical Semantics 15 (2024) 19.

[40] P. Mateiu, A. Groza, Ontology engineering with large language models, in: 2023 25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), IEEE, 2023, pp. 226–229.

[41] M. S. Abolhasani, R. Pan, Ontokgen: A genuine ontology and knowledge graph generator using large language model, in: 2025 Annual Reliability and Maintainability Symposium (RAMS), IEEE, 2025, pp. 1–6.

[42] J. Vrolijk, V. Poslavsky, T. Bijl, M. Popov, R. Mahdavi, M. Shokri, Ontology learning for esco: Leveraging llms to navigate labor dynamics, Proceedings of the 2nd workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM 2024) (2023).

[43] A. M. Bran, A. Oarga, M. Hart, M. Lederbauer, P. Schwaller, Ontology-retrieval augmented generation for scientific discovery, Under review as a conference paper at ICLR 2025 (2025).

[44] H. Yang, L. Xiao, R. Zhu, Z. Liu, J. Chen, An llm supported approach to ontology and knowledge graph construction, in: 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2024, pp. 5240–5246.

[45] A. Mukanova, M. Milosz, A. Dauletkaliyeva, A. Nazyrova, G. Yelibayeva, D. Kuzin, L. Kussepova, Llm-powered natural language text processing for ontology enrichment (2024).

[46] M. Val-Calvo, M. E. Aranguren, J. Mulero-Hernández, G. Almagro-Hernández, P. Deshmukh, J. A. Bernabé-Díaz, P. Espinoza-Arias, J. L. Sánchez-Fernández, J. Mueller, J. T. Fernández-Breis, Ontogenix: Leveraging large language models for enhanced ontology engineering from datasets, Information Processing & Management 62 (2025) 104042.

[47] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An llm supported approach to ontology and knowledge graph construction, arXiv preprint arXiv:2403.08345 (2024).

[48] B. Zhang, V. A. Carriero, K. Schreiberhuber, S. Tsaneva, L. S. González, J. Kim, J. de Berardinis, Ontochat: a framework for conversational ontology engineering using language models, in: European Semantic Web Conference, Springer, 2024, pp. 102–121.

[49] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146. URL: https://aclanthology.org/Q17-1010/. doi:10.1162/tacl_a_00051.

[50] C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, B. Dhoedt, Learning semantic similarity for very short texts, in: 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1229–1234. doi:10.1109/ICDMW.2015.86.

KG-Quizzer: Refinamento de Prompts via Grafos de Conhecimento para a Geração Automática de Questionários em Português

Davisson Medeiros[1], Gabriel Leite[1], André Gomes Regino[2], Victor Jesus Sotelo Chico[1], Ferrucio de Franco Rosa[2] and Julio Cesar dos Reis[1]

[1] *Instituto de Computação, Universidade Estadual de Campinas, UNICAMP, Brazil*
[2] *Center for Technology Information Renato Archer, Brazil*

Resumo

A falta de recursos educacionais acessíveis e personalizados continua sendo uma barreira significativa para uma educação de qualidade no Brasil. Este estudo investiga como as tecnologias da Web Semântica, combinadas com Modelos de Linguagem de Grande Porte (LLMs), podem facilitar a geração automatizada de questionários educacionais, reduzindo assim a carga de trabalho dos professores. Propomos um framework que integra Grafos de Conhecimento (KGs) e técnicas de Engenharia de Prompts para aprimorar a qualidade das questões geradas. Nossa pesquisa avalia o impacto do uso de triplas RDF extraídas de KGs, comparando a injeção de prompts em formatos brutos e verbalizados, bem como o papel do Few-shot Learning na melhoria da eficácia dos LLMs na tarefa investigada. Os resultados experimentais indicam que triplas verbalizadas melhoram a clareza linguística, enquanto triplas RDF brutas aprimoram a estrutura e a precisão factual. Constatamos que a contextualização de prompts por meio do Few-shot Learning aumenta significativamente a coerência e relevância das questões geradas. Nosso estudo destaca o valor de combinar conhecimento estruturado com modelos generativos para aplicações educacionais baseadas em conhecimento.

Abstract

The lack of accessible and personalized educational resources remains a significant barrier to quality education in Brazil. This study investigates how Semantic Web technologies, combined with Large Language Models (LLMs), can facilitate the automated generation of educational questionnaires, thereby reducing teacher workload. We propose a framework that integrates Knowledge Graphs (KGs) and Prompt Engineering techniques to enhance the quality of the generated questions. Our research evaluates the impact of using RDF triples extracted from KGs, comparing prompt injection from raw and verbalized formats, as well as the role of Few-shot Learning in improving LLM effectiveness in the investigated task. Experimental results indicate that verbalized triples enhance linguistic clarity, while raw RDF triples improve structure and factual accuracy. We found that prompt contextualization via Few-shot Learning significantly boosts the coherence and relevance of the generated questions. Our study highlights the value of combining structured knowledge with generative models for knowledge-based educational applications.

Keywords

Knowledge Graphs, Text Generative Models, Verbalization, Questionnaire Generation, LLMs

1. Introdução

A falta de acesso à educação de qualidade no Brasil é um desafio social crítico, afetando principalmente professores e alunos de escolas públicas em regiões de baixa renda. Obstáculos como a superlotação das salas e a falta de materiais didáticos adequados dificultam o ensino personalizado e podem comprometer o desempenho acadêmico dos estudantes [1]. Nesse cenário, questionários surgem como uma ferramenta

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

dvsmedeiros.research@gmail.com (D. Medeiros); gabriel.dfleite@gmail.com (G. Leite); aregino@cti.gov.br (A. G. Regino); v265173@dac.unicamp.br (V. J. S. Chico); ferrucio.rosa@cti.gov.br (F. d. F. Rosa); jreis@ic.unicamp.br (J. C. d. Reis)
0009-0004-6522-1039 (D. Medeiros); 0009-0003-3302-0673 (G. Leite); 0000-0001-9814-1482 (A. G. Regino); 0000-0001-9245-8753 (V. J. S. Chico); 0000-0001-9504-496X (F. d. F. Rosa); 0000-0002-9545-2098 (J. C. d. Reis)

valiosa para avaliação contínua e formativa [2], mas sua elaboração e correção manual demandam um tempo e esforço que sobrecarregam ainda mais os professores, tornando sua implementação um desafio.

A criação automática de questionários eficazes apresenta desafios, como a garantia de alinhamento com os objetivos pretendidos, a formulação clara de perguntas e a adaptação às diversas necessidades dos respondentes [3]. Esses desafios são potencializados pela crescente demanda por personalização no ensino e relevância das avaliações, que precisam ser adaptadas a diferentes níveis de conhecimento [4].

Nesse cenário, o problema central consiste em como desenvolver métodos e ferramentas computacionais de apoio na construção de avaliações educacionais, em específico, questionários, que sejam efetivos, personalizáveis e adaptáveis aos diferentes níveis de conhecimento dos alunos, tendo potencial de contribuir para diminuir a sobrecarga de trabalho dos professores. O processo de criação manual de questionários se demonstra oneroso, dificultando o processo de acompanhamento contínuo e individualizado dos alunos.

Esta pesquisa objetiva especificar, desenvolver e avaliar um framework computacional (**KG-Quizzer**) para a geração automática de questionários (pares de pergunta e resposta). Assumimos que ao combinar dados estruturados provenientes de Grafos de Conhecimentos (KGs) pode beneficiar o refinamento de prompts em Modelos de Linguagem de Grande Escala (LLMs). Neste trabalho, o termo KGs é empregado para se referir a bases de conhecimento representadas por triplas RDF, com foco específico na DBpedia. Nossa solução visa efetuar uma integração dessas tecnologias, pois os KGs atuam como conhecimento externo estruturado sobre o tema das questões a serem geradas. Em nossa solução, KGs beneficiam como meio de minimizar a alucinação (fenômeno no qual o texto gerado pelo LLM contêm imprecisões ou não faz sentido) dos modelos LLM e melhor contextualizar o tema na geração do questionário.

Mais especificamente, visamos verificar qual a influência da utilização de triplas RDFs, extraídas de KGs, na qualidade da geração das perguntas e respostas por LLMs, assim como, qual a diferença de efetividade (qualidade dos resultados) ao fornecer as triplas em formato RDF bruto em comparação com a sua forma verbalizada em linguagem natural. Verificamos igualmente qual o ganho de qualidade dos resultados obtidos ao se empregar a técnica de *Few-shot Learning* [5] para guiar o LLM na geração de perguntas e respostas.

Os resultados experimentais revelam uma forte dependência da qualidade da geração em relação às estratégias de tratamento de dados e ao modelo de linguagem utilizado. Evidenciou-se que a aplicação de técnicas de engenharia de *prompt*, como a verbalização de triplas de conhecimento, melhora significativamente a legibilidade e a simplicidade do texto gerado, priorizando a qualidade linguística. A utilização das triplas em formato RDF resulta em textos mais bem estruturados, priorizando a corretude. Constatou-se que a qualidade textual é aprimorada pela contextualização do *prompt*. Dentre as abordagens avaliadas, a técnica *Few-shot Learning* se mostrou como a mais efetiva, alcançando um padrão de qualidade superior na geração dos questionários. Esses achados reforçam que uma estratégia de interação bem definida é fundamental para a geração de conteúdo confiável e de alto valor.

A contribuição central desta investigação materializa-se na proposição do nosso framework para a construção de questionários educacionais, fundamentado na sinergia entre KG e LLM. Nosso estudo oferece contribuições específicas ao investigar sistematicamente o impacto que o uso de triplas de conhecimento, provindas de um KG, exerce sobre a qualidade da geração de questionários. Nossa pesquisa contribui ao avaliar comparativamente os ganhos de se utilizar as triplas em seu formato RDF bruto em contraste com uma representação verbalizada. Avançamos em se analisar originalmente o impacto da técnica *Few-shot Learning* na efetividade da produção dos questionários em português, buscando otimizar a interação com o LLM e a relevância das perguntas geradas.

O restante desse artigo está estruturado da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 apresenta o Framework proposto; a Seção 4 descreve a metodologia experimental; a Seção 5 descreve os resultados; a Seção 6 discute nossos achados. A Seção 7 conclui o artigo.

2. Trabalhos Relacionados

A aplicação de LLMs para gerar e avaliar conteúdo educacional tornou-se uma área de pesquisa proeminente. Analisamos trabalhos recentes que exploram o uso de LLMs na Geração Automática de Questões e no desenvolvimento de novos paradigmas de avaliação.

He *et al.* [6] analisaram o uso do modelo de linguagem ChatGLM na geração de questões para o currículo de tecnologia da informação no ensino médio. Compararam questões geradas por humanos e pela IA, avaliadas por especialistas segundo critérios como relevância, dificuldade e imparcialidade. O ChatGLM apresentou desempenho comparável ao dos humanos na maioria dos critérios, evidenciando potencial para reduzir a carga de trabalho docente. Contudo, demonstrou limitações na formulação de questões que exigem aplicação prática, revelando desafios persistentes dos LLMs em traduzir conhecimento teórico para contextos reais.

Kido *et al.* [7] investigaram a geração de questões de múltipla escolha para o Exame Nacional de Enfermagem Japonês, com foco na criação de distratores, identificada como a parte mais desafiadora. Os autores utilizaram dados de exames anteriores e avaliaram quatro LLMs: GPT-4, ChatGPT (utilizando as técnicas de *Fine-Tuning* e *Few-Shot*) e o *Japanese Stable LM* (JSLM). Efetuaram o ajuste fino de ChatGPT e JSLM com somente 193 questões, enquanto GPT-4 e ChatGPT usaram aprendizado em contexto (Few-Shot). Introduziram novas métricas de avaliação com base em similaridade semântica, complementadas por avaliação humana. O ChatGPT ajustado teve melhor efetividade em precisão e recuperação, e o GPT-4 gerou distratores mais aceitos por especialistas, apesar de menos correspondentes ao conjunto de referência. As métricas de similaridade mostraram-se mais adequadas que as tradicionais para avaliar distratores com equivalência semântica. Os resultados indicaram que o ajuste fino pode ser mais eficaz que *Few-shot Learning* em tarefas específicas de Geração Automática de Questões (GAQ), mesmo com conjuntos de dados pequenos.

Karvinen [8] propuseram um sistema que usa o GPT-3.5-Turbo para gerar pré-questões de múltipla escolha com base em livros didáticos, voltado a alunos do ensino superior. Para lidar com a limitação de tokens dos LLMs e evitar alucinações, o sistema segmenta o conteúdo dos livros e aplica busca por similaridade (usando FAISS com embeddings da OpenAI) para selecionar partes relevantes conforme o termo de busca. As questões geradas foram avaliadas manualmente quanto à fundamentação no material, legibilidade, variedade e falhas. O sistema teve melhor desempenho que o ChatGPT em relação à fundamentação (94,9% vs. 72,6%) e gerou questões com legibilidade adequada. A arquitetura baseada em segmentação e recuperação de contexto se demonstrou eficaz para garantir precisão factual. O estudo destacou o uso das questões como ferramenta pedagógica para engajamento e aprendizagem, ampliando o papel dos LLMs na educação para além da avaliação.

Chico *et al.* [9] propuseram uma estrutura baseada em IA generativa para a criação automática de quizzes de múltipla escolha (MCQs) a partir de textos em linguagem natural em português, com foco em contextos educacionais. A abordagem considerou modelos de linguagem do tipo encoder-decoder, como o T5 e sua variante ajustada para o português (PTT5), combinando técnicas de *fine-tuning* e engenharia de prompts. Para avaliar a qualidade das questões geradas, exploraram métricas automáticas de legibilidade, complexidade sintática e diversidade lexical (Brunet, Yngve), além de análises qualitativas. Os resultados indicaram que modelos refinados, especialmente o PTT5, apresentaram melhor desempenho em legibilidade e diversidade em comparação às variantes do Flan-T5. Revelaram que a engenharia de prompts, embora menos custosa computacionalmente, gerou resultados comparáveis ao *fine-tuning* em diversos cenários. A proposta amplia o uso de LLMs na educação, não somente como suporte à avaliação, mas também como ferramenta ativa na geração de atividades que promovem o engajamento e o aprendizado dos estudantes.

Nossa abordagem, denominada *KG-Quizzer*, concentra-se na definição de um pipeline para a geração de pares de pergunta e resposta em português via LLMs, fundamentados em conhecimento estruturado via KGs. Utilizamos LLMs para gerar e validar os pares de pergunta e resposta, tendo como fonte de conhecimento KGs e técnicas de *engenharia de prompt*, como *Few-shot.* O foco na automação da criação de conteúdo avaliativo alinha-se com He *et al.* [6], que também visam reduzir o esforço humano via assistência das LLMs. Semelhante a Kido *et al.* [7] e Karvinen [8], o *KG-Quizzer* envolve a definição e a geração de questões a partir de uma base de conhecimento específica, no nosso caso a *DBpedia*.

A literatura aborda desafios centrais, como garantir a fidelidade factual do conteúdo gerado e a dificuldade dos modelos em criar questões que exijam aplicação prática do conhecimento. Adicionalmente, a avaliação da qualidade e da imparcialidade dos LLMs emerge como um campo crítico, uma vez que um questionário mal formulado representa um risco ao aprendizado do aluno. Embora trabalhos relacionados (por exemplo, He *et al.*[6], Kido *et al.*[7], Karvinen [8]) se concentrem em gerar questões a partir de fontes de texto em linguagem natural, nossa abordagem é específica para investigar sistematicamente como a qualidade da geração é impactada pela integração de conhecimento estruturado de KGs. Diferentemente de Karvinen [8], que utiliza LLMs para geração baseada na recuperação de trechos de texto, nossa proposta se concentra na injeção de triplas RDF ou verbalizadas extraídas de KGs, permitindo um controle mais rigoroso da proveniência factual.

Adicionalmente, enquanto Kido *et al.* [7] exploraram métricas automatizadas e He *et al.*[6] dependem de especialistas humanos para avaliação, nossa abordagem emprega uma LLM como validador em um pipeline de duas etapas, inspirado no conceito de *LLM-as-a-Judge* e a utilização de métricas textuais Nilcmetrix [10], para uma filtragem automática. Isso distingue nosso trabalho de He *et al.* [6]e se alinha mais com a ênfase na sinergia homem-máquina de Kido *et al.*[7] para otimizar a qualidade da geração.

Uma distinção fundamental é que, enquanto os trabalhos de Kido *et al.* [7] e Karvinen [8] operam sobre fontes de dados textuais, nosso escopo atual atua na integração e análise do impacto direto dos KGs. Embora nossa abordagem compartilhe estratégias baseadas em LLM com outros trabalhos, nossa investigação se distingue por seu foco na análise comparativa do uso de conhecimento estruturado (triplas RDF vs. verbalizações) e técnicas de aprendizado em contexto para a tarefa de geração de questionários. Apesar do considerável número de estudos sobre o uso de LLMs no suporte à criação de conteúdo educacional, a literatura carece de pesquisas sobre o refinamento da geração de questionários através da injeção controlada de conhecimento estruturado de KGs no contexto de Língua Portuguesa.

3. KG-Quizzer

KG-Quizzer é um framework desenvolvido para automatizar a geração de pares pergunta-resposta em linguagem natural na língua portuguesa, combinando a capacidade generativa dos LLMs com dados estruturados extraídos de KGs. A proposta central do framework é enriquecer (refinar) semanticamente os *prompts* utilizados na geração, por meio da incorporação controlada de conhecimento factual e exemplos, visando a produção de questionários mais relevantes, coerentes e semanticamente consistentes.

A Figura 1 apresenta o funcionamento do framework com um diagrama de arquitetura conceitual. O diagrama apresenta os módulos principais do framework, suas entradas e saídas, descrevendo o fluxo completo desde a entrada de um tópico, tema do questionário, até a validação final dos pares gerados.

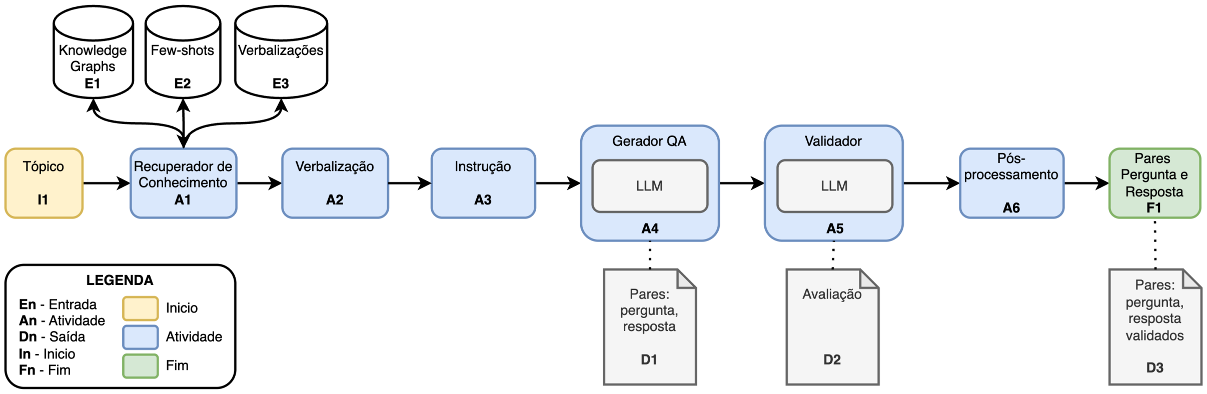


Figura 1: Arquitetura Conceitual do KG-Quizzer.

O processo se inicia a partir de um tópico de entrada (I1), que representa o conceito central a ser explorado. Esse tópico é utilizado como chave para recuperação de informação estruturada, processadas pelo Módulo (A1), que reúne dados provenientes de três fontes principais: triplas extraídas de KGs por meio de consultas[1] [2]SPARQL, como a consulta de abstracts e a consulta de triplas relacionadas) ($E_1 = \{\tau_1, \tau_2, \ldots, \tau_n\}$, onde cada τ_i é uma tripla RDF extraída do grafo de conhecimento), exemplos de pares pergunta-resposta representativos do domínio selecionado pelo tópico ($E_2 = \{(q_1, a_1), (q_2, a_2), \ldots, (q_m, a_m)\}$, onde q_i é uma pergunta e a_i sua respectiva resposta) e versões verbalizadas das triplas ($E_3 = \{v_1, v_2, \ldots, v_n\}$, onde cada v_i é uma sentença em linguagem natural correspondente à tripla τ_i).

As triplas obtidas podem ser utilizadas diretamente ou processadas pelo Módulo (A2), responsável por sua verbalização. Esse processo transforma representações formais (como RDF) em sentenças em linguagem natural, facilitando a integração com os modelos de linguagem. O Módulo (A3) realiza a montagem da instrução (*prompt*) [3] [4] [5], organizando os elementos disponíveis: tópico, contexto estruturado, exemplos, fatos e instruções, conforme o cenário configurado, utilizando três variantes principais de prompts: um zero-shot, outro zero-shot com triplas, e um few-shot com triplas e exemplos. A construção da instrução é flexível e parametrizável, permitindo diferentes estratégias de contextualização e orientação da geração.

O *prompt* estruturado é enviado ao Módulo (A4), responsável pela geração dos pares pergunta-resposta com apoio de um modelo de linguagem. A saída é armazenada como artefato intermediário ($D_1 = \{(q'_1, a'_1), (q'_2, a'_2), \ldots, (q'_k, a'_k)\}$, em que cada par é produzido por um modelo de linguagem).

Na sequência, os pares são submetidos ao Módulo (A5), que realiza uma validação automática via um segundo modelo de linguagem. Essa etapa visa julgar a qualidade das perguntas e respostas quanto à completude, clareza, consistência, adequação linguística e correção gramatical, gerando um artefato de avaliação ($D_2 = \{r_1, r_2, \ldots, r_k\}$, em que r_i representa o resultado da avaliação do par (q'_i, a'_i)).

A última etapa do processo é conduzida pelo Módulo (A6), que executa o pós-processamento dos pares gerados. Isso inclui tarefas como normalização textual, cálculo de métricas como as fornecidas pelo pacote NILC-Metrix [10], capazes de refletir a coesão, coerência e nível de complexidade textual, filtragem de pares malformados e organização dos resultados. O produto dessa etapa (F1) é um conjunto consolidado e validado de pares pergunta-resposta ($D_3 = \{(q''_1, a''_1), \ldots, (q''_j, a''_j)\}$).

4. Metodologia da Avaliação Experimental

A arquitetura conceitual de KG-Quizzer foi instanciada em um protocolo experimental sistemático, ilustrado na Figura 2, no qual cada componente da arquitetura foi operacionalizado ao longo das etapas de ponta a ponta.O experimento foi conduzido visando mensurar o impacto de diferentes estratégias de enriquecimento semântico na qualidade dos pares pergunta-resposta produzidos pelo KG-Quizzer.

4.1. Protocolo Experimental

O Protocolo Experimental para avaliação do KG-Quizzer (Figura 2) representa a instanciação da arquitetura do KG-Quizzer no experimento conduzido. Cada módulo foi operacionalizado com artefatos concretos e fontes reais de dados, seguindo a mesma estrutura modular apresentada na Figura 1.

Os tópicos (E1) foram extraídos da versão em português do Stanford Question Answering Dataset (SQuAD) [11], o qual consiste de um dataset de compreensão de leitura com perguntas sobre artigos da Wikipédia, onde a resposta para cada pergunta é um trecho de texto extraído da própria passagem de leitura, e enviados ao módulo de consulta (A1), que coleta: a) dados estruturados da DBpedia [12] (E2), b) exemplos *Few-shot* do próprio SQuAD (E3) e; c) as verbalizações disponíveis para predicados RDF (E4). Esses dados são processados para retornar triplas RDF e gerar suas versões verbalizadas (A2).

[1] https://github.com/dvsmedeiros/kg-quizzer/blob/main/queries/abstract.txt

[2] https://github.com/dvsmedeiros/kg-quizzer/blob/main/queries/triplas-relacionadas.txt

[3] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/zero-shot.txt

[4] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/zero-shot-triplas.txt

[5] https://github.com/dvsmedeiros/kg-quizzer/tree/main/prompts/few-shot-triplas.txt

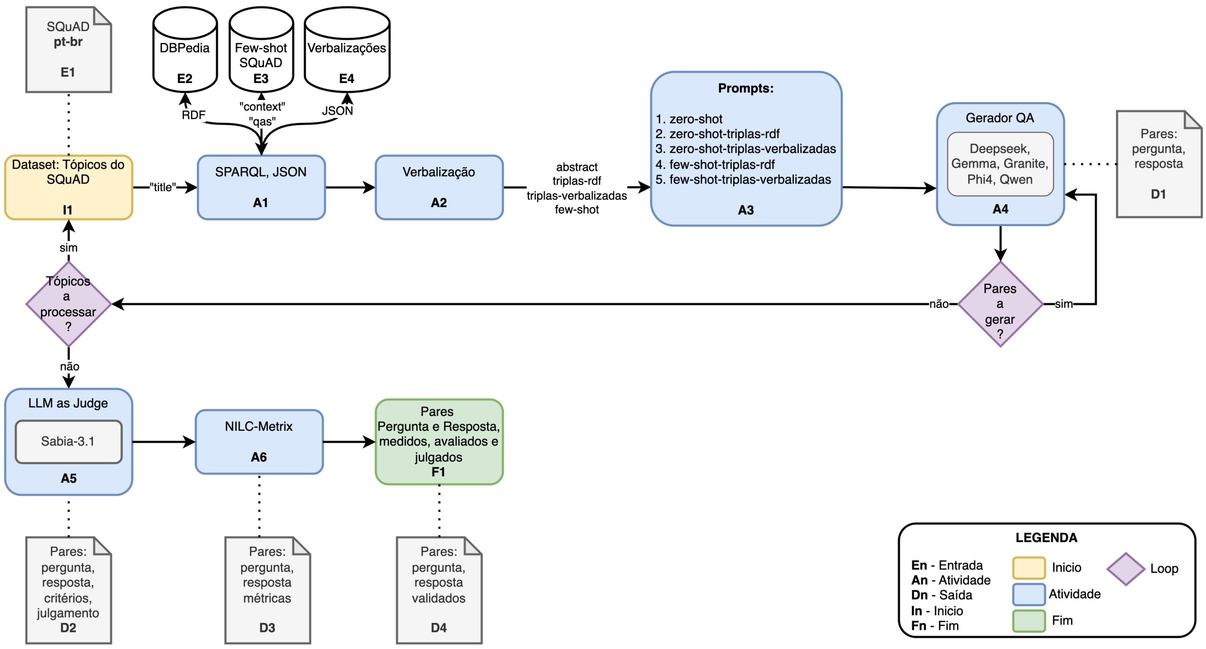


Figura 2: Protocolo Experimental para Avaliação do KG-Quizzer.

Com os dados processados, os prompts são compostos (A3) conforme cinco estratégias experimentais distintas, que combinam a presença ou ausência de exemplos (*Zero-shot* ou *Few-shot*) com diferentes formas de expressão do conhecimento (RDF ou verbalizada). A Tabela 1 apresenta um resumo conceitual desses cinco cenários de geração.

Tabela 1
Descrição dos cenários experimentais para geração de questões. T_i refere-se a um dos $i = 35$ tópicos e M_j a um dos $j = 18$ modelos utilizados. Foram definidos cinco tipos de cenário: C0 corresponde ao zero-shot simples, enquanto C1 a C4 envolvem triplas RDF em diferentes combinações de técnica e representação. Cada conjunto com triplas (zero-shot ou few-shot) soma 1.260 cenários considerando ambos os formatos (verbalizado e RDF), totalizando 3.150 combinações conforme: $T \times M + 2 \times T \times (F \times M)$.

Identificador	Tópico	Modelo	Técnica de Prompt	Representação da Tripla	Nº de Cenários
C0	T_i	M_j	Sem contexto adicional	-	630
C1	T_i	M_j	Presente (Few-Shot)	Verbalizada	630
C2	T_i	M_j	Presente (Few-Shot)	RDF (Não Verbalizada)	630
C3	T_i	M_j	Ausente (Zero-Shot)	Verbalizada	630
C4	T_i	M_j	Ausente (Zero-Shot)	RDF (Não Verbalizada)	630

A aplicação sistemática desses cinco cenários a todos os 35 tópicos e 18 modelos de linguagem resultou em um total de 3.150 combinações experimentais. Apresentamos a decomposição dessa quantidade com base nos tipos de estratégias e formatos empregados.

4.2. Seleção dos Tópicos e Consulta ao Grafo de Conhecimento

Os tópicos utilizados no experimento foram selecionados a partir da versão em português do conjunto de dados *SQuAD*, amplamente reconhecido por sua variedade temática e relevância educacional. Todos os 35 tópicos disponíveis foram incorporados, garantindo diversidade e equilíbrio entre diferentes domínios do conhecimento. A seleção contempla áreas como ciência, história, geografia e tecnologia (*e.g.*, “Sistema Imunitário”, “Peste Negra” e “Teoria da Complexidade Computacional”). Cada tópico foi utilizado como base para a recuperação de conhecimento estruturado na DBpedia e também para a seleção de exemplos *Few-shot*.

4.3. Representações Semânticas e Construção de Prompts

O protocolo experimental foi elaborado para avaliar de forma sistemática o impacto de diferentes estratégias de *prompt* e representações semânticas na qualidade dos pares pergunta-resposta gerados pelo framework. A construção dos prompts considerou três fatores principais: (a) o tipo de estratégia de geração (*zero-shot* ou *few-shot*), (b) a forma de expressão do conhecimento (sem triplas, com triplas RDF ou verbalizadas); e (c) a presença ou não de exemplos. Essas combinações resultaram em cinco cenários experimentais: zero-shot sem conhecimento adicional, zero-shot com triplas RDF, zero-shot com triplas verbalizadas, few-shot com triplas RDF e few-shot com triplas verbalizadas.

As triplas RDF foram extraídas da *DBpedia* em inglês, devido à maior cobertura e disponibilidade de resultados. Após a recuperação, foram aplicados filtros para remover predicados semanticamente irrelevantes, como metadados técnicos, links externos ou descritores genéricos. O predicado `dbo:abstract` também foi excluído, pois seu conteúdo é consultado separadamente e incorporado ao contexto textual. Para facilitar a interpretação pelos modelos, as triplas RDF foram verbalizadas em linguagem natural[6]. Cada predicado foi mapeado manualmente para expressões equivalentes em português, adotando como critério de tradução a preservação do sentido semântico original. Por exemplo, `dbo:capital` foi representado como "tem como capital" ou "*has as capital*". Uma tripla como *`Brazil`* `dbo:capital Brasília` resultou em uma sentença mais legível: "O Brasil tem como capital Brasília".

Nos cenários com triplas, a quantidade incluída no *prompt* foi ajustada com base na capacidade de contexto dos modelos de linguagem. Embora alguns suportem até 128.000 tokens, o limite foi fixado em 8.000 tokens para compatibilidade geral. Parte desse espaço foi reservada para instruções, resumo do tópico e formatação; o restante foi utilizado para triplas. Se a quantidade de triplas excedesse o espaço disponível, aplicava-se um corte sequencial conforme a ordem da consulta SPARQL. Os exemplos *Few-shot* foram obtidos a partir do SQuAD em português, considerando somente aqueles relacionados ao mesmo tópico e com perguntas válidas e respondidas. O número máximo por prompt foi de dez exemplos, valor alinhado ao número de pares avaliados posteriormente, assegurando consistência entre geração e avaliação.

4.4. LLMs Utilizados e Configuração de Geração

Foram selecionados 18 LLMs com base em critérios que asseguram diversidade arquitetural, adequação à tarefa e viabilidade de execução local. Os modelos abrangem diferentes famílias, como Deepseek, Gemma, Granite, Phi4 e Qwen, permitindo comparações entre arquiteturas distintas e evitando viés associado a uma única arquitetura. Para contemplar diferentes capacidades, os modelos foram categorizados por porte: pequeno (até 4 bilhões de parâmetros), médio (entre 5B e 14B) e grande (entre 15B e 32B).

Somente modelos com janelas de contexto de pelo menos 8.000 tokens foram incluídos, garantindo que *prompts* extensos, compostos por conhecimento factual e exemplos, fossem processados integralmente. Por outro lado, modelos com mais de 32B de parâmetros foram descartados devido a limitações práticas de execução local e custo computacional. Todos os modelos selecionados estavam disponíveis para execução em ambiente controlado (sem necessidade de uso de API externa), favorecendo a reprodutibilidade e o controle sobre o ambiente experimental. A Tabela **??** apresenta os LLMs utilizados, com suas respectivas famílias, tamanhos (porte) e janelas de contexto.

A geração dos pares pergunta-resposta foi conduzida com um perfil padronizado de hiperparâmetros, definido a partir de experimentos preliminares: *temperature* de 0.7, *top-k* de 40 e *top-p* de 0.95. Essa configuração buscou um equilíbrio entre coerência, diversidade lexical e fluência nas respostas geradas.

[6] https://github.com/dvsmedeiros/kg-quizzer/blob/main/resources/predicados_verbalizados_pt_en.csv

4.5. Avaliação dos Pares Gerados

A avaliação da qualidade dos pares pergunta-resposta gerados foi conduzida por meio de duas abordagens complementares: (a) análise "quantitativa", baseada em métricas linguísticas; e (b) avaliação "qualitativa", automatizada por modelo de linguagem (*LLM-as-a-Judge*). Essa combinação permitiu analisar os pares tanto sob aspectos formais quanto sob critérios semânticos e linguísticos.

Na avaliação quantitativa, utilizamos as métricas fornecidas pelo pacote *NILC-Metrix* [10], com foco em complexidade textual e diversidade lexical. As principais métricas foram simple_word_ratio [13], brunet [14] e Yngve [15]. A Métrica `brunet` [14] mede a legibilidade, no qual que valores maiores indicam maior acessibilidade textual. A métrica `Yngve` [15] avalia a complexidade sintática das sentenças, na qual valores mais altos indicam um texto mais complexo e difícil.. A Métrica `simple_word_ratio` [13] representa a proporção de palavras simples, refletindo a facilidade de leitura. Esses indicadores foram utilizados para verificar se os pares gerados apresentam linguagem acessível, comparável a questionários educacionais destinados a públicos diversos. Os valores de referência utilizados têm como base estudos prévios como [5] e [16].

A avaliação qualitativa foi conduzida com o modelo *Sabia-3*, especializado na língua portuguesa. Esse modelo é de uma família distinta dos usados na tarefa de geração. Cada par pergunta-resposta foi avaliado com base em cinco dimensões: i) *Complexidade* - considera se o par exige raciocínio, síntese ou inferência, indo além da simples memorização, e se possui uma estrutura válida de pergunta e resposta; ii) *Completude* - avalia se a resposta abrange adequadamente todos os aspectos solicitados na pergunta, evitando omissões relevantes; iii) *Corretude* - analisa se a resposta está de fato correta e coerente com a pergunta e o contexto, verificando a ausência de contradições ou inconsistências; iv) *Fluidez* - examina se o par apresenta clareza na leitura e estrutura linguística natural em português, sem trechos truncados ou confusos; e v) *Qualidade* - observa se há correção gramatical e ortográfica no texto, garantindo que a linguagem esteja adequada considerando um falante nativo de português. Cada dimensão foi pontuada pelo modelo em uma escala de 1 a 5, em que 1 representa desempenho insatisfatório e 5 indica excelência. O modelo atribui notas individualizadas para cada dimensão e calcula uma média final por par. Ele também foi instruído [7] a produzir uma justificativa textual que explica os motivos das avaliações atribuídas, promovendo transparência e interpretabilidade ao processo de julgamento.

O julgamento foi realizado com base em uma execução que gerou 6.300 arquivos, correspondentes a 3.150 cenários. Para cada cenário, foi produzido um arquivo de texto contendo a resposta do modelo e um arquivo em formato de valores separados por vírgula (CSV) com os pares extraídos. Casos não extraídos foram automaticamente tratados manualmente ou com apoio de LLMs para recuperação dos dados. Após o processamento, 3.137 cenários foram considerados válidos para avaliação, representando 99,59% do total. Os demais foram descartados por ausência de pares, falhas de extração ou problemas de formatação. Um total de 31.737 pares pergunta-resposta foi avaliado, uma vez que alguns cenários produziram mais do que os 10 pares solicitados.

5. Resultados Experimentais

Esta seção apresenta os resultados experimentais. A Subsecao 5.1 reporta a análise quantitativa e qualitativa com *LLM-as-a-Judge* da qualidade das gerações. A Subseção 5.2 relata uma seleção de exemplos positivos e negativos. Resultados completos e análises suplementares, como a de custo computacional[8] estão disponíveis no repositório[9].

5.1. Resultados da Avaliação dos Pares

A Tabela 2 apresenta que ambas as estratégias (*RDF e Verbalização*) de triplas afetam positivamente a construção de pares pergunta-resposta, como demonstrado nos resultados de todas as métricas em

[7] https://github.com/dvsmedeiros/kg-quizzer/blob/main/prompts/llm-as-judge.txt

[8] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/analise_de_custo_computacional.png

[9] https://github.com/dvsmedeiros/kg-quizzer

relação ao cenário sem triplas (*None*). Observamos que a utilização de triplas RDF resulta em pares mais bem estruturados, refletido pelos valores superiores de Complexidade, Completude e Corretude. A estratégia de triplas verbalizadas resulta em pares com melhor qualidade linguística, refletido pelos valores superiores de fluência e qualidade de português.

Tabela 2
Resultados da Análise de Estratégias de tratamento das triplas aplicando *LLM-as-a-Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
None	2.450	3.500	4.119	4.033	4.273	3.675
RDF	**2.550**	**3.705**	**4.327**	4.163	4.365	**3.822**
Verbalizado	2.497	3.674	4.317	**4.175**	**4.403**	3.813
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 3 apresenta que as respostas produzidas utilizando a estratégia de triplas verbalizadas possuem menor complexidade sintática (porém, menor legibilidade), evidenciado pelos valores de `A-yngve` e `A-brunet` comparados às outras abordagens. Além disso, a estratégia de verbalização também se sobressai em relação ao *BeQuizzer* [9], sugerindo que um modelo de linguagem geral, quando alimentado com dados bem pré-processados (neste caso, verbalizados), pode superar ferramentas especializadas na criação de texto acessível.

Tabela 3
Resultados da Análise de Estratégias de tratamento das triplas relativo às métricas do Nilcmetrix [10].

	Q-yngve	**Q-brunet**	**Q-swr**	**A-yngve**	**A-brunet**	**A-swr**
None	2.214	4.813	0.374	1.588	3.872	0.221
RDF	2.292	4.853	0.331	1.660	3.698	0.181
Verbalizado	2.280	4.869	0.342	**1.275**	**3.077**	0.148
BeQuizzer	2.396	5.202	-	1.142	3.476	-
Adapt2Kids	2.48	11.03	0.74	2.48	11.03	0.74
Leg2Kids	1.60	12.87	0.76	1.60	12.87	0.76
Total	2.271	4.851	0.344	1.491	3.482	0.176

A Tabela 4 apresenta que a utilização de contexto afetou positivamente na construção de pares pergunta-resposta, como mostrado através de valores de métricas superiores quando inserido contexto. Nota-se que o cenário que gera a melhor efetividade em todas as métricas é a utilização de *Few-shot*, reforçando o impacto positivo dessa técnica.

Tabela 4
Resultados da Análise de Estratégias de Contextualização aplicando *LLM-as-a-Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
Sem Contexto	2.450	3.500	4.119	4.033	4.273	3.675
Zero Shot	2.460	3.631	4.289	4.152	4.333	3.773
Few Shot	**2.586**	**3.748**	**4.355**	**4.186**	**4.435**	**3.862**
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 5 apresenta que as respostas produzidas utilizando *Few-shot* possuem menor complexidade sintática, porém, com menor legibilidade, evidenciado pelos valores de `A-yngve` e `A-brunet` comparados às outras abordagens. As métricas para as Perguntas (Q-) mostram menos variação entre as estratégias abordadas, sugerindo que o principal impacto da estratégia de contextualização estaria na qualidade da resposta gerada, e não na pergunta.

Tabela 5
Resultados da Análise de Estratégias de Contextualização relativo às métricas do Nilcmetrix [10].

	Q-yngve	Q-brunet	Q-swr	A-yngve	A-brunet	A-swr
Sem Contexto	2.214	4.813	0.374	1.588	3.872	0.221
Zero Shot	2.309	4.856	0.317	1.687	3.724	0.147
Few Shot	2.262	4.865	0.356	**1.248**	**3.052**	0.183
BeQuizzer	2.396	5.202	-	1.142	3.476	-
Adapt2Kids	2.48	11.03	0.74	2.48	11.03	0.74
Leg2Kids	1.60	12.87	0.76	1.60	12.87	0.76
Total	2.271	4.851	0.344	1.491	3.482	0.176

A Tabela 6 apresenta que a família de modelos qwen3 demonstra os melhores resultados, com destaque para o modelo qwen3:32b. Há uma tendência de modelos maiores obterem resultados melhores, mas a arquitetura do modelo é mais decisiva que o tamanho do modelo, segundo os nossos experimentos.

Tabela 6
Resultados da Análise Comparativa entre Modelos de Linguagem aplicando *LLM as a Judge*. A_1 = Complexidade, A_2 = Completude, A_3 = Corretude, A_4 = Fluência, A_5 = Qualidade do Português e A_6 = Média de Avaliação. Os modelos estão ordenados relativo à Média de Avaliação.

	A_1	A_2	A_3	A_4	A_5	A_6
qwen3:32b	2.532	4.040	4.602	4.305	4.586	4.013
qwen3:14b	2.565	3.942	4.551	4.271	4.535	3.973
qwen3:8b	2.562	3.843	4.498	4.265	4.533	3.940
gemma3:27b	2.483	3.904	4.543	4.237	4.506	3.935
phi4:14b	2.584	3.874	4.487	4.215	4.478	3.928
qwen3:30b	2.517	3.820	4.497	4.241	4.527	3.921
deepseek-r1:32b	2.556	3.841	4.492	4.257	4.427	3.914
qwen2.5:14b	2.525	3.851	4.461	4.205	4.501	3.909
qwen2.5:32b	2.461	3.817	4.458	4.233	4.505	3.895
qwen3:4b	2.536	3.766	4.425	4.225	4.473	3.885
deepseek-r1:14b	2.529	3.780	4.437	4.242	4.344	3.867
granite3.1-dense:8b	2.561	3.725	4.327	4.172	4.444	3.846
gemma3:4b	2.523	3.576	4.304	4.227	4.478	3.822
deepseek-r1:8b	2.593	3.658	4.299	4.165	4.314	3.806
gemma3:12b	2.446	3.575	4.321	4.165	4.450	3.791
qwen2.5:3b	2.478	3.376	4.009	4.081	4.302	3.649
deepseek-r1:7b	2.430	2.855	3.396	3.750	3.854	3.257
deepseek-r1:1.5b	2.245	2.350	2.803	3.205	3.143	2.749
Total	2.509	3.652	4.282	4.142	4.362	3.789

A Tabela 7 mostra que os modelos que têm uma menor complexidade sintática e menor legibilidade (evidenciado pelos valores de A-yngve e A-brunet) são os pertencentes às famílias gemma e qwen. Mais especificamente, os modelos gemma3:27b e qwen3:32b, como pode ser visto no repositório[10].

5.2. Exemplos de Resultado Positivo e Negativo

A Figura 3 apresenta um exemplo de resultado com as melhores estratégias identificadas na Seção 5.1, utilização de triplas em formato RDF e *Few-shot* com o modelo qwen3:32b. Nota-se que o resultado exibe uma pergunta conceitualmente rica e uma resposta tecnicamente precisa. Isso justifica sua avaliação “quase perfeita” e valida os altos scores do modelo em corretude e fluência. Um exemplo de

[10] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/nilcmetrix_por_modelo.png

Tabela 7
Resultados da Análise Comparativa entre Modelos de Linguagem relativo às métricas do Nilcmetrix [10].

	Q-yngve	**Q-brunet**	**Q-swr**	**A-yngve**	**A-brunet**	**A-swr**
DeepSeek	2.313	4.781	0.298	1.737	3.778	0.176
Gemma	2.273	4.878	0.393	1.287	3.110	0.182
Mistral	2.236	4.854	0.348	1.722	4.231	0.198
Phi	2.355	5.102	0.335	1.460	3.459	0.161
Qwen	2.241	4.855	0.354	1.391	3.349	0.173
Total	2.271	4.851	0.344	1.491	3.482	0.176

resultado com as piores estratégias é mostrado no repositório [11].

Pergunta: Por que o número 1 não é considerado primo?

Resposta: Para preservar a unicidade da fatoração em números primos no teorema fundamental da aritmética.

Critérios de Avaliação
- Complexidade: 4
- Completude: 5

Corretude: 5
Fluência: 5
Qualidade do Português: 5

Média de Avaliação: 4.8

Justificativa: A pergunta exige uma compreensão conceitual sobre números primos e a importância do teorema fundamental da aritmética, o que confere uma complexidade elevada. A resposta é completa, correta e precisa, abordando exatamente o motivo pelo qual o número 1 não é considerado primo. A fluidez e a qualidade do português são excelentes, sem erros gramaticais ou ortográficos e com uma estrutura clara e natural.

Figura 3: Par pergunta-resposta gerado e avaliação realizada pelo *LLM-as-a-Judge* utilizando a estratégia de triplas RDF, com *Few-shot* e modelo `qwen3:32b` para o tópico "Número primo".

6. Discussão

A utilização de triplas RDF para incrementar o contexto do *prompt* resultou em textos com melhor estrutura linguística, estando conforme a influência originalmente esperada da utilização de triplas RDF para incrementar o *prompt*, visto que se trata de um conhecimento estruturado.

Relativo à utilização de triplas verbalizadas, esperava-se que os textos gerados tivessem um comportamento linguístico mais próximo do humano, visto que esse é o objetivo da verbalização. Os resultados se evidenciaram em concordância com o esperado, através da melhora textual, na qualidade linguística e na facilitação da legibilidade dos textos gerados.

A melhora na efetividade dos resultados foi guiada pela hipótese de que os resultados se tornassem progressivamente melhores à medida que o contexto no *prompt* fosse enriquecido. A análise dos resultados demonstrou que essa expectativa se concretizou empiricamente. Partindo de uma base com somente o tópico, em que as respostas são frequentemente genéricas e "imprevisíveis", a introdução da técnica *Zero-shot* representa um salto de qualidade significativo ao fornecer um objetivo explícito, direcionando o modelo sobre o que fazer. Os melhores resultados observados foram com a abordagem *Few-shot*, que além de fornecer o objetivo, incorpora exemplos práticos diretamente no *prompt*. Essa transição — de um comando vago para uma instrução explícita e, por fim, para uma tarefa demonstrada com exemplos — representa um caminho claro para maximizar a precisão, o controle e a confiabilidade das gerações textuais na tarefa estudada.

Com relação aos diversos modelos *open-source* utilizados, dois pontos merecem destaque. Primeiramente, confirmamos a expectativa de que modelos com maior número de parâmetros geram textos de

[11] https://github.com/dvsmedeiros/kg-quizzer/blob/main/anexo/caso_qualitativo_ruim.png

qualidade superior, conforme observado na Tabela 6. O fator mais notável foi o impacto preponderante da arquitetura do modelo no resultado. Isso se torna evidente com os modelos da família qwen, que, mesmo possuindo casos de menor número de parâmetros, apresentaram uma efetividade superior ao de outros modelos teoricamente mais potentes.

Os resultados através do framework *KG-Quizzer* indicaram que, embora ele consiga produzir perguntas com uma estrutura sintática adequada e relevante para crianças de 4 a 11 anos, refletido pelos valores de *yngve*, o vocabulário utilizado é excessivamente complexo. As métricas de legibilidade, *brunet* e `simple_word_ratio`, apontaram que a dificuldade de leitura das perguntas é de duas a três vezes superior ao nível considerado ideal para essa faixa etária. Essa complexidade textual pode se tornar um obstáculo para potenciais alunos, sobretudo para as crianças com menor proficiência de leitura.

O framework proposto é suficientemente genérico para ser adaptado a diferentes domínios, idiomas e modelos, mantendo a separação entre os componentes de recuperação, construção de *prompt*, geração e avaliação. Essa modularidade favorece tanto extensões quanto sua aplicação em diferentes contextos.

Expandir esta investigação para incluir a geração de distratores (alternativas incorretas à questão) é um passo relevante para viabilizar a construção completa de questionários de múltipla escolha. Distratores bem elaborados aumentam a complexidade das questões e permitem avaliar com mais precisão o conhecimento dos alunos, promovendo um aprendizado mais desafiador. Essa extensão pode utilizar métricas de redes complexas para identificar nós semanticamente semelhantes às perguntas, gerando distratores plausíveis. Esse é um caminho de pesquisa futura.

7. Conclusão

A geração automática com qualidade de perguntas e respostas em Português para fins educativos ainda é um desafio em aberto. Este estudo propôs e avaliou um framework que combina dados estruturados via KGs com *prompts* na geração de questionários. Nossos resultados demonstraram que a utilização de triplas no formato RDF se apresentou mais efetiva quando o foco é a estrutura textual. Verbalizar as triplas se apresentou útil quando se deseja qualidade linguística e facilidade de leitura dos questionários produzidos, ou seja, mais próxima da linguagem humana. Constatou-se a melhora progressiva na geração de perguntas através do enriquecimento de contexto. A utilização de *Zero-shot* mostrou-se mais efetiva que a *baseline* (sem contexto enriquecido), e *Few-shot* apresentou os melhores resultados. Como trabalhos futuros, planejamos a simplificação do vocabulário e a adaptação dos *prompts* para reduzir a complexidade textual. Visamos ainda explorar diferentes abordagens de verbalização e outros KG públicos, como Wikidata e YAGO, assim como desenvolver um KG próprio para o framework proposto. Visamos ainda definir e implementar métricas mais refinadas e específicas para validar a coerência semântica.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4 and Gemini in order to assist with the following tasks, as defined in the CEUR-WS GenAI Usage Taxonomy: “Paraphrase and reword”, “Improve writing style”, and “Grammar and spelling check”. These contributions were limited to the revision and refinement of existing content. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication’s content.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil, projeto #301337/2025-0. Assim como o apoio fornecido pela BTG Pactual e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

[1] M. H. D. SILVA, Déficit de aprendizagem causado pela super lotação na sala de aula, Amazon Live Journal v.6 (2024) 1–15. URL: https://zenodo.org/records/13230100. doi:10.5281/zenodo.13230100.

[2] C. Kivunja, Why Students Don't Like Assessment and How to Change Their Perceptions in 21st Century Pedagogies, Creative Education 6 (2015) 2117–2126. URL: https://www.scirp.org/journal/paperinformation?paperid=61373. doi:10.4236/ce.2015.620215, number: 20 Publisher: Scientific Research Publishing.

[3] G. Kurdi, J. Leo, B. Parsia, U. Sattler, S. Al-Emari, A Systematic Review of Automatic Question Generation for Educational Purposes, Int J Artif Intell Educ 30 (2020) 121–204. URL: https://doi.org/10.1007/s40593-019-00186-y. doi:10.1007/s40593-019-00186-y.

[4] B. P. Solis Trujillo, D. Velarde-Camaqui, C. A. Gonzales Nuñez, E. V. Castillo Silva, M. d. P. Gonzalez Said de la Oliva, The current landscape of formative assessment and feedback in graduate studies: a systematic literature review, Front. Educ. 10 (2025). URL: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1509983/full. doi:10.3389/feduc.2025.1509983, publisher: Frontiers.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[6] L. He, Y. Chen, X. Hu, Application of large language models in automated question generation: A case study on chatglm's structured questions for national teacher certification exams, arXiv preprint arXiv:2408.09982 (2024).

[7] Y. Kido, H. Yamada, T. Tokunaga, R. Kimura, Y. Miura, Y. Sakyo, N. Hayashi, Automatic question generation for the japanese national nursing examination using large language models., in: CSEDU (1), 2024, pp. 821–829.

[8] L. Karvinen, Using a large language model-based system to generate pre-questions in a quiz format for students focused in self-directed learning, Master's thesis, Itä-Suomen yliopisto, 2023.

[9] V. J. S. Chico, J. F. Tessler, R. Bonacin, J. C. dos Reis, Bequizzer: Ai-based quiz automatic generation in the portuguese language, in: A. Rapp, L. Di Caro, F. Meziane, V. Sugumaran (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2024, pp. 237–248.

[10] S. E. Leal, M. S. Duran, C. E. Scarton, N. S. Hartmann, S. M. Aluísio, NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese, Lang Resources & Evaluation 58 (2024) 73–110. URL: https://doi.org/10.1007/s10579-023-09693-w. doi:10.1007/s10579-023-09693-w.

[11] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 2383–2392. doi:10.18653/v1/D16-1264.

[12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, Z. Ives, Dbpedia: A nucleus for a web of open data, in: international semantic web conference, Springer, 2007, pp. 722–735.

[13] M. T. C. Biderman, Dicionário Didático de Português, Editora Ática, São Paulo, 1998.

[14] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, E. Asp, Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech, in: IEEE International Conference Mechatronics and Automation, 2005, volume 3, 2005, pp. 1569–1574 Vol. 3. URL: https://ieeexplore.ieee.org/document/1626789. doi:10.1109/ICMA.2005.1626789, iSSN: 2152-744X.

[15] V. H. Yngve, A model and an hypothesis for language structure, Proceedings of the American Philosophical Society 104 (1960) 444–466. URL: http://www.jstor.org/stable/985230.

[16] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, The curious case of neural text degeneration, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=rygGQyrFvH.

Notes on the use of the powertype pattern*

André M. Demori[1,*,†], Julio Cesar Cardoso Tesolin[2,†] and Maria Cláudia Reis Cavalcanti[3,†]

[1]*Military Institute of Engineering, Praça Gen. Tibúrcio, 80 - Urca, Rio de Janeiro - RJ, 22290-270*

Abstract

Identifying the need for multi-level modeling patterns in conceptual models is a non-trivial task, especially when such needs are implicit or are heavily context dependent. This paper proposes some guidelines for identifying the use of the Powertype pattern, grounded in the Multi-Level Theory (MLT), using domain-independent competency questions to support this identification. The approach is illustrated through a case study that requires multi-level representations for entities and relationships at the type level. Our proposal aids the detection of hidden structures in models, contributing both to the construction of multi-level conceptual models and of ontology-driven conceptual models.

Keywords

modeling patterns, Powertype pattern, multi-level modeling, multi-level theory

1. Introduction

During the process of building a conceptual model, some entities present aspects and behaviors that lead to representing them as both class and an individual. While the former presents predicative characteristics, the latter presents specific characteristics. Fonseca et al. [1] state that this phenomenon occurs mainly when classes are also subject to classification, commonly seen in biological taxonomies.

This conceptual modeling challenge can be dealt with by using multi-level modeling patterns. A well-known multi-level pattern is the *Powertype* pattern [2, 3], which occurs when instances of one type (powertype) are specializations of another type (basetype). However, recognizing the need to use this modeling pattern is not straightforward. It requires the modeler to identify hidden entities and relationships that are not immediately apparent when designing the first versions of a conceptual model.

This work proposes a set of steps (guidelines) to support the modeler in the use of multi-level modeling patterns and how they can improve the resulting conceptual model. We analyze specific situations where the *Powertype* pattern may be applied and the benefits it brings. Additionally, for each situation, we formulate useful domain-independent competency questions that cut across modeling levels, which would otherwise remain unanswered. Furthermore, uncovering implicit model entities and relationships, discovering the possible patterns applicable to provide semantic enrichment helps in restoring the ontological commitment. In this manner, we consider these guidelines a relevant contribution of this paper to the process of ontological analysis of conceptual models because of its didactic nature on showing the multi-level structures potential to enrich representations. Finally, to illustrate the use of the proposed guidelines, a case study is presented, in which we revise the conceptual modeling of an existing ontology by applying the *Powertype* pattern[1]. Additionally, we show how the punning technique can be used to implement the identified *powertypes*.

This article is structured as follows: Section 2 provides the theoretical background on multi-level

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

*Corresponding author.

†These authors contributed equally.

andredemori@ime.eb.br (A. M. Demori); jcctesolin@ime.eb.br (J. C. C. Tesolin); yoko@ime.eb.br (M. C. R. Cavalcanti)

0000-0002-0533-3395 (A. M. Demori); 0000-0002-0240-4506 (J. C. C. Tesolin); 0000-0003-4965-9941 (M. C. R. Cavalcanti)

[1] We use *Powertype* with a capital letter, followed by pattern, when referring to the Powertype pattern. We use *powertype* with a lowercase letter, when referring to the conceptual element itself (a type of type), used to denote classes that classify other classes.

conceptual modeling and foundational ontologies. Section 3 reviews the main related works that address multi-level modeling and the *Powertype* pattern. Section 4 presents the core contributions of this work. Section 5 illustrates the proposed approach through a case study, demonstrating both the revised modeling of an ontology and its operationalization. Finally, Section 6 summarizes the main findings and outlines directions for future research.

2. Background

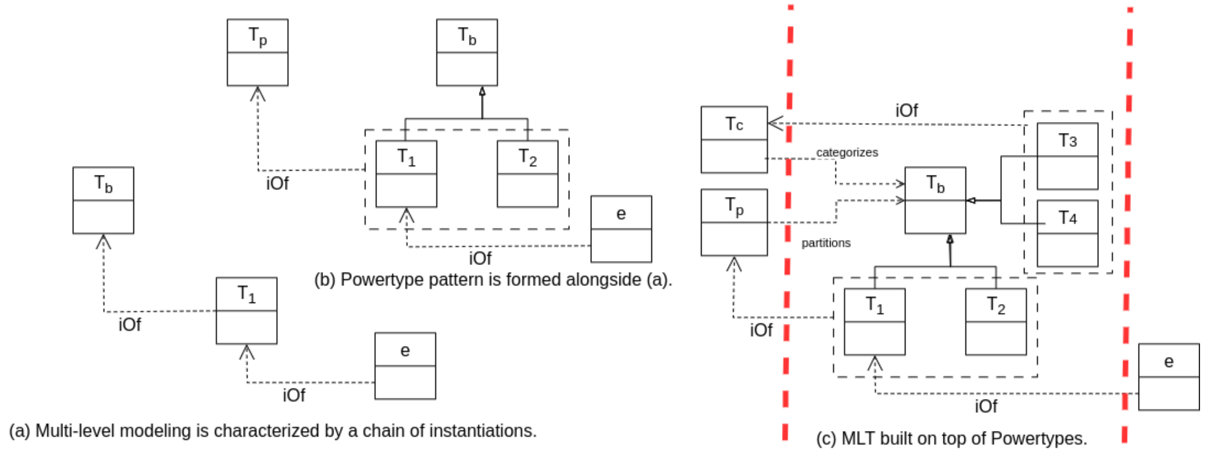
Foundational ontologies such as UFO [4], GFO [5], BFO, and DOLCE [6, 7] provide formal semantics that are crucial for conceptual modeling. They enable the expression of both explicit and implicit characteristics of entities in a domain, supporting semantic enrichment and a strongly axiomatized representation. Leveraging these semantics requires an ontological analysis to uncover the metaphysical and structural foundations of the entities and their interrelations.

Ontology-driven conceptual modeling (ODCM) involves systematically analyzing a domain—often starting from pre-existing structures such as database schemas or ER diagrams—to reveal its underlying ontological commitments, identifying and addressing modeling patterns [8], including multi-level patterns [9] and *truthmakers* [10]. It is a critical challenge, as these are often absent or only implicitly represented in existing models. Reification is one of the strategies that could be useful in this process, once it brings to light hidden entities, promoting them to first-class citizens [10].

2.1. Multi-Level Conceptual Modeling

According to [11], an intension[2] characterizes each type. This intension identifies whether the type applies to a particular entity. Thus, if the intension of a type T_1 applies to an entity *e*, then *e* is said to be an instance of T_1. However, there are modeling cases where an intension of a type T_p applies to type T_1, i.e., type T_1 is an instance of a type T_p. This chain is illustrated in Figure 1(a). The *Powertype* pattern in conceptual modeling appears alongside the occurrence of this chain of instantiations, as shown in 1(b).

MLT incorporates two notions of powertype, one based on Cardelli's [3] definition and another based on Odell's [2]. According to Cardelli [3], this pattern is defined as follows: if T_b is a type, then T_p is the type whose instances are all the subtypes of T_b; if T_1 is an instance of T_p then T_1 is a subtype of T_b. T_p is referred to as *powertype(T_b)*, and T_b is referred to as *basetype*. On the other hand, Odell considers that there are specializations of the base type that may not be an instance of a *powertype*. MLT shows how both definitions are related to each other and how they can be used in different manners [11].



(b) Powertype pattern is formed alongside (a).

(a) Multi-level modeling is characterized by a chain of instantiations.

(c) MLT built on top of Powertypes.

Figure 1: Multi-level modeling evolution.

[2] In this work we use the word intension meaning the set of characteristics or properties by which the referent or referents of a given word are determined. https://www.collinsdictionary.com/dictionary/english/intension (accessed September 9, 2025).

To create a well-founded theory about multi-level modeling, Carvalho et al. developed the Multi-Level Theory (MLT) [11, 12, 13, 14], which primarily characterizes the orders of types and the partitioning and categorization relations between types. Figure 1(c) summarizes in a visual way, the contribution of MLT to the *Powertype* pattern.

By the theory's definition, types that have *Individuals* as instances are classified as first-order types (1stOT), and types whose instances are first-order types are classified as second-order types (2ndOT), and so on, thus creating chains of instantiations. These are the basic types of Multi-Level Theory. Once the *Powertype* pattern has been recognized, it is necessary to check whether there is a partitioning or a categorization relationship between *powertype* and *basetype*. According to MLT, a *powertype* partitions the *basetype* if and only if each instance of the *basetype* is an instance of exactly one instance of the *powertype*.

Also, according to MLT [11], the categorization relation occurs between a *powertype* and a *basetype* when the intension of the *powertype* defines that its instances specialize the *basetype* according to a specific classification criterion. A variation of the categorization relation is called Complete categorization, in which the classification criteria defined by the intension of the *powertype* guarantees that each instance of the *basetype* is an instance of at least one instance of the *powertype*. A third variation is called Disjoint categorization, in which each instance of the *basetype* is an instance of at most one instance of the *powertype*.

3. Related Works

The academic literature has been proposing several techniques and theories to deal with multi-level modeling. Neumayr et al. [15] addressed the issue of multiple levels in conceptual modeling through what is called Multiple Level Objects (m-objects) and Multiple Level Relationships (m-relationships), which provide a representation with multiple levels of abstraction, encapsulating the different levels that relate to a single domain concept and integrating aspects of the different semantic abstraction hierarchies (aggregation, generalization and classification) into a single concretization hierarchy. However, as Carvalho et al. [13] noted in their work, this abstraction of levels can render the modeler unable to express whether instances of a higher-order type are disjoint and/or comprehensive types and also unable to determine the meta-properties.

Guizzardi and Almeida discuss stability patterns in conceptual models [16]. Then, the authors analyze several stability patterns, including multi-level modeling and reification of intrinsic and relational aspects. The authors provide an interesting explanation of how reifying intrinsic aspects in quality spaces and relations leads to more stable conceptual models. Moreover, the higher-order types can help to create conceptual models that focus on invariant aspects, turning a less rigid conceptual model.

In turn, Lara, Guerra, and Cuadrado propose a deep explanation about when and how to use multi-level modeling [17], discussing situations where the use of multi-level modeling is beneficial for the representation. The authors also introduce different techniques to work with multi-level and also address discussions about multi-level patterns, including the Powertype pattern. Halpin also presents a didactic work to describe some challenges in modeling subtypes [18], examining several subtype issues such as derivation options, subtype rigidity, and subtype migration. However, the author focuses on ORM technique to explore these concepts and not on a well-founded theory of a foundational ontology.

Guizzardi et al. [9] proposed a study towards an ontological analysis of powertypes, analyzing issues such as (i) powertype instances as universals, (ii) powertype instances as mereological sums (iii) powertype instances as variable embodiments. Also, the authors discussed the identity of instances and the classification relation (isClassifiedBy).

Despite these works offering similar approaches to address or analyze multi-level issues in conceptual modeling, they do not develop domain-independent competence questions, which helps reinforce that this pattern should be applied. Moreover, this paper proposes a didactic approach for working with multi-level in conceptual models, which can be considered a complement to related works.

4. Multi-Level Modeling Guidelines

As already mentioned, the problem we want to address is the difficulty in applying multi-level modeling patterns while modeling a domain. In this direction, this section presents some guidelines for applying multi-level modeling, assuming that it may bring benefits. Section 4.1 introduces the first steps based on the use of the reification technique. Section 4.2 focuses on identifying the appropriate level of attributes in a multi-level model. Finally, Section 4.3 takes us to additional steps where we deal with related hierarchical restrictions on multiple levels.

4.1. Multi-Level Identification First Steps

As mentioned in subsection 2.1, a typical multi-level modeling case is when a chain of instantiations occurs, i.e., when an entity e is an instance of T_1, which in turn is an instance of T_p. If we can evolve into a richer multi-level modeling, like the one in Figure 1(c), then it becomes possible to answer the following set of Competency Questions (CQ):

CQ1: Which types categorize T_b?
CQ2: Which types partition T_b?
CQ3: In which subtypes can one classify (categorize) an instance of T_b according to T_c?
CQ4: In which subtypes can one partition instances of T_b according to T_p?

However, it is not an easy task to identify this pattern and also to predict such CQs. It is a usual modeling practice to include types that have attributes or relationships that classify their instances. For example, consider the following two domain models: one that has a type T_b, which is classified by an $attr_1$ (Figure 2(a)), and another one where type T_b is classified by another type T_p (Figure 2(b)). In both cases, the above competency questions could not be answered. In the first case (Figure 2(a)), it is not clear what the meaning of $attr_1$ to T_b is. Similarly, in the second case (Figure 2(b)), the meaning of the "isClassifiedBy" relationship is unclear, as it does not distinguish whether the classification partitions or categorizes T_b, unless the cardinality of the relationship is explicitly defined. In other words, there are hidden entities and relationships that should come to light.

Guarino et al. [10] use reification in conceptual modeling to identify *truth-making* patterns. It involves making explicit entities while modeling a domain that would otherwise be implicit. Similarly, we examine the use of reification to make explicit the *Powertype* pattern. Thus, inspired by Guarino et al., we defined a set of steps that depart from reifying attributes and evolve the model up to the point that it forms a multi-level pattern. After applying these steps, it becomes possible to answer the previously defined CQ. It is worth mentioning that these steps are one of the possible methods to obtain a rich multi-level modeling.

Figure 2 illustrates the first steps to identify the *Powertype* pattern. It begins with a simple type T_b and its attribute $attr_1$ (Figure 2(a)). If $attr_1$ is a classifying attribute for T_b that is essential for its characterization, and which needs to have its own attributes, then it must be reified into a type T_p. Thus, in **Step 1**, the model evolves to Figure 2(b), where $attr_1$ corresponds to T_p, which is connected to T_b through the *isClassifiedBy* relationship.

Next, in **Step 2**, we need to check if there are specializations of type T_b that need to be represented. For instance, there might be other relationships departing from T_b that indicate that a subset of it may be related to a T_x, as shown in Figure 2(c). This indicates that a specialization of T_b may be hidden and must be revealed [19]. In this case, the model evolves to Figure 2(d), where a *subtype* set is created for type T_b.

Then, in **Step 3**, for each revealed specialization, we check if the subtype extension matches the extension of type T_p. If this is the case, it means that the *Powertype* pattern is formed, as shown in Figure 2**(e)**, where a dotted line represents the *instance Of* (iOf) relationship between type T_p(*powertype*) and the *subtypes* (T_1 and T_2). In addition, the relationship between the *basetype* and the *powertype* in this case is (*partitions*), which means T_b is completely partitioned into disjoint subtypes according to T_p.

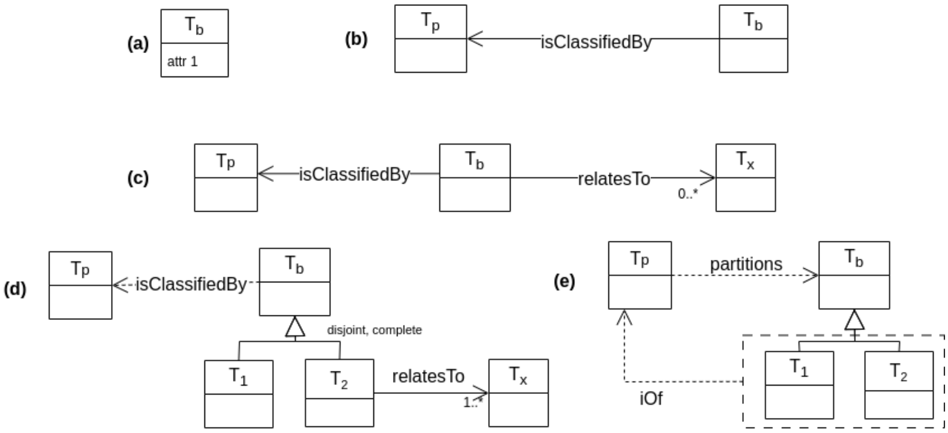


Figure 2: Multi-level identification first steps.

To illustrate the steps described above, Figure 3 presents a domain-specific example, where the *Powertype* pattern is formed having as its *basetype* the type Person. Initially, type Person has been modeled as having two classifying attributes, namely, *Employee Position* and *Academic Role* (Figure 3(a)). An instance of type Person, "Maria", would be represented with values "Coordinator" and "Professor", respectively, assigned to those attributes. After applying the steps (Figure 3(b)-(e)), the type Person becomes a *basetype*, which is categorized by the *Academic Role* type, partitioned by the *Employee Position* type, and specializes in two *subtypes*. Based on this example, we can formulate and answer the previously defined CQs, as follows:

- Which types categorize *Person*? Academic Role
- Which types partition *Person*? Employee Position
- In which Subtypes can I classify (categorize) an instance of *Person* according to *Academic Role*? Student and Professor
- Which *subtypes* of *Person* are instances of Employee Position? Manager, Coordinator, and Analyst.

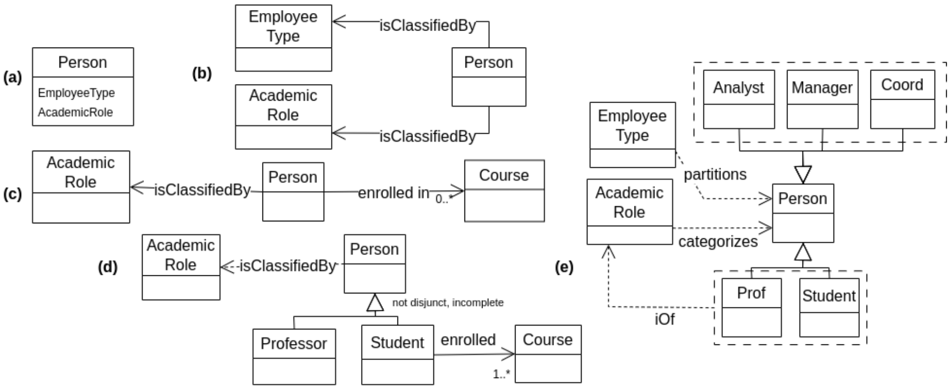


Figure 3: First steps applied to an example.

4.2. Modeling Type-Level Attributes

In addition to the initial steps described in Section 4.1, we present other cases in which the *Powertype* pattern can improve modeling by representing attributes in higher-order types.It is not rare to find models where the *basetype* and the *powertype* are collapsed into a single type, mixing their attributes. A typical modeling case is presented in Figure 4(a), where all instances of T_1 are assigned the same value for attribute $regAttr3$, while a different value is assigned to it in all instances of T_2. This kind of

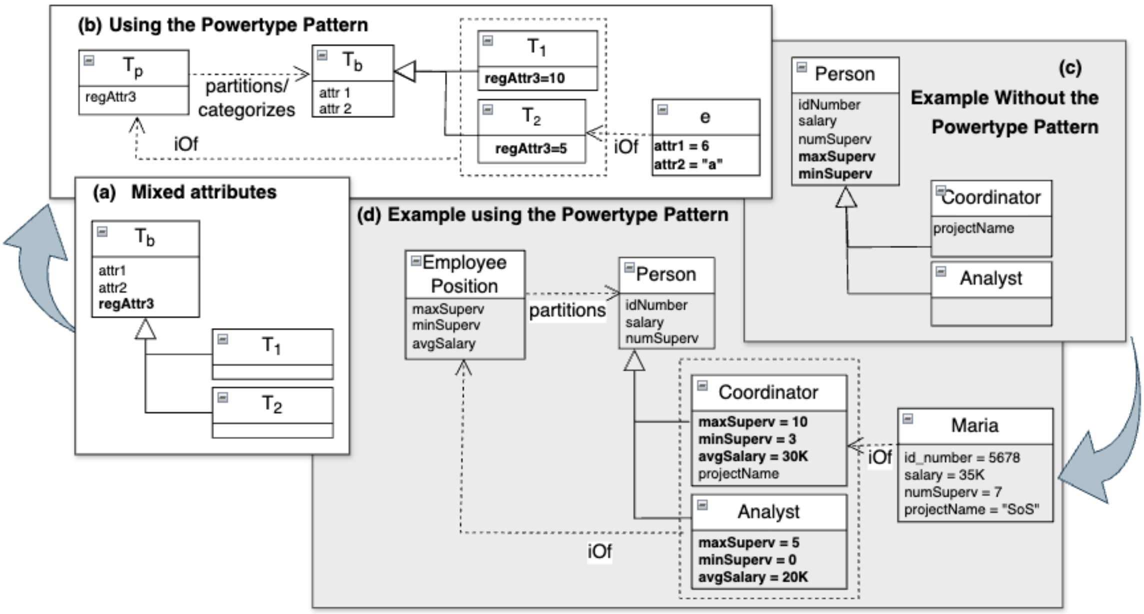


Figure 4: Modeling type-level attributes.

attributes have been called *regularity attributes* [9]. They aim at capturing regularities over lower-level type instances (i.e., instances of T_b subtypes), and constraining them, acting as patterns or defining intervals that must be obeyed.

Thus, we propose **Step 4**, as a good practice in this case: create the *powertype* T_p, as shown in Figure 4(b), and move the $regAttr3$ attribute to T_p. Thus, this attribute will be assigned values in instances of T_p (T_1 and T_2), constraining instances of T_1 and T_2 (e). Also, it may be useful to express patterns or rules that instances of T_b must comply with, e.g. $T_b.attr1 > T_p.regAttr3$. With this modeling step, these constraints are made explicit and allow us to formulate the following CQ:

CQ5: What are the value restrictions that an attribute $attr$ of a type T_i should respect for all instances $e \in T_i$?

Figure 4(c) illustrates this step using a domain-specific model. In this example, we have the type *Person* corresponding to T_b from Figure 4(a). *Person* subtypes share attributes such as *minSuperv* and *maxSuperv*, which are regularity attributes. Indeed, in all instances of the type $Coordinator$, the $minSuperv$ attribute will be assigned 3, while in all instances of the type $Analyst$, it would be assigned 0. After identifying these regularity attributes, the type *Employee Position* (*powertype(Person)*) is created, and, the regularity attributes are moved to the new type, as shown in Figure 4(d). With this, type $Person$ instances are now constrained according to the *Employee Position* type. Moreover, *Person.numSuperv* attribute values must obey the type *Employee Position*. In the example, "Maria" is an instance of *Person* whose *numSuperv* attribute is set to 7, respecting the *Coordinator* type interval constraint ($maxSuperv = 10 \wedge minSuperv = 3$).

Finally, some attributes may be derived from the instances of T_b, which are known as *resultant* attributes [9]. These attributes are usually numeric and calculated based on the subtype instances (e.g. average, rate, percentage, etc.), and thus they should belong to the higher-order type T_p. In the example of Figures 4(c)-(d), note that type *Person* has *salary* attribute that will have values for all its instances. This may be an opportunity to create the *resultant* attribute *avgSalary* at the *Employee Position* type, and maintain its values based on the values of the *salary* attribute.

After applying this step, it becomes possible to formulate and answer the following CQs:

- What is the average salary of a Coordinator?

- What is the minimum and maximum number of supervisions that are required to become a Coordinator?

4.3. Modeling Subtype Restrictions

In addition to the steps outlined in the previous sections, when modeling multiple levels, the modeler should be aware of the relationships that connect subtypes. It might be useful to represent these relations at the *Powertype* level. Suppose, for instance, that there is a relationship named *connectsTo* between T_b subtypes, T_1 and T_2, as shown in Figure 5(a), meaning that it connects instances of T_1 to instances of T_2. However, once the *Powertype* T_p has subtypes of T_b as its instances, this rule can be expressed by a self-relationship named *mayConnectTo* at the *Powertype*, as shown in Figure 5(a). Thus, **Step 5** consists of expressing the subtype relationships as a rule at a higher-order level type (T_p and its self-relationship), which allows the following CQ to be formulated and answered:

CQ6: Which instances of T_b could be connected to each other? In other words, given an instance of T_b, to which other instances could it be connected?

Furthermore, let us consider that the *basetype* T_b may have many subtypes, forming a flat tree with many siblings in a sequence. Now, suppose the *connectsTo* relationship occurs between many of these sibling pairs (e.g. (T_1,T_2), (T_2,T_3),...,(T_{n-1},T_n)). Although these relationships have similar semantics, in principle, it is not possible to generalize them as each one connects instances of specific subtype pairs. However, once the connection rule is preserved through the relationship *mayConnectTo* at type T_p, it becomes possible to generalize those relationships, substituting them by a single self-relationship at T_b. Therefore, with the help of the *Powertype* pattern, **Step 6** simplifies the model by reducing the number of relationships, as shown in Figure 5(b).

Figure 5(c) shows a domain-specific example where it is necessary to represent relationships between two pairs of subtypes in the hierarchical structure. The *analystRespondsTo* relationship specifically connects instances of types *Analyst* and *Coordinator*, while the *coordinatorRespondsTo* connects instances of types *Coordinator* and *Manager*.

Applying the steps described before, the model evolves to the one in Figure 5(d). It leverages the *Powertype* pattern to represent only two relationships, a self-relationship in the *basetype Person* and another in the *powertype Employee Position*. With this remodeling, it becomes possible to answer the following CQ: *Which instances of type Person can respond to another?* In other words, which restrictions should be applied to the instantiation of the *basetypeRespondsTo* relationship between *Person* instances? According to the *powertype Employee Position* and its instances, the answer is: *Analysts* can respond to *Coordinators*, and *Coordinators* can respond to *Managers*. It is worth noting that steps 5 and 6 can be applied recursively at each level of a higher hierarchical structure.

5. Reviewing The MiScOn Ontology

This case study addresses multi-level modeling challenges in the development of the MiScOn ontology (Military Scenario Ontology) [20], in light of the steps and competency questions discussed in Section 4. MiScOn captures both tactical and communication system aspects of military operations, grounded in military doctrines and validated by domain experts. It was chosen because it was developed based on the SABiO methodology [21], which distinguishes clear stages of the development process and their respective output artifacts. First, a reference ontology is generated and, later, the corresponding operational version is generated.

Moreover, OntoUML was used as MiScOn modeling language. While reviewing the MiScOn reference artifact, this allowed the modeler not only to ground the modeling on the Unified Foundational Ontology (UFO) [22], but also to count on the support of the Visual Paradigm OntoUML plugin[3], which was used to generate the MiScOn operational artifact.

[3] https://github.com/OntoUML/ontouml-vp-plugin

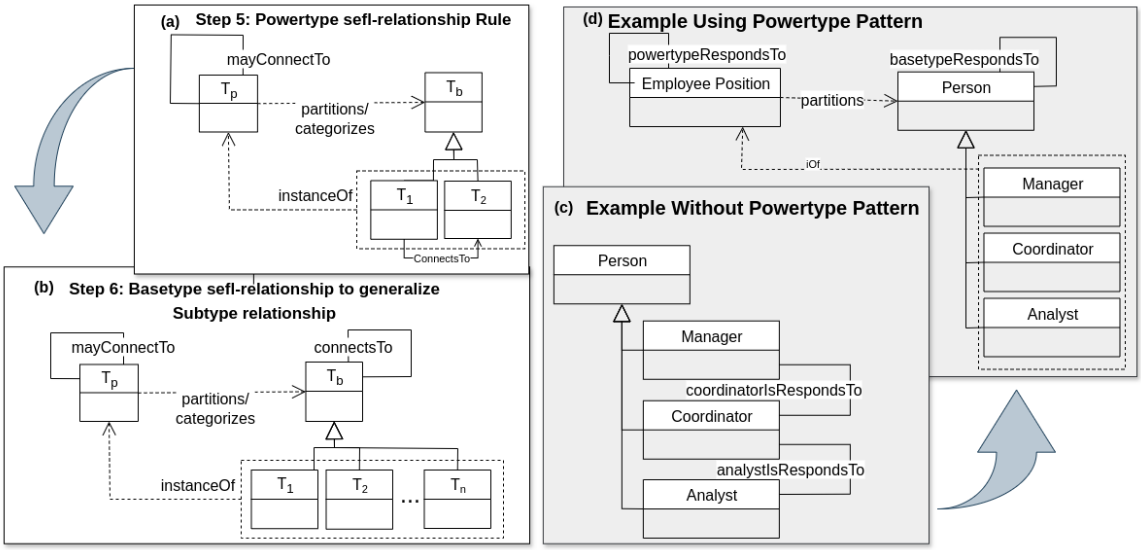


Figure 5: Modeling subtypes restrictions through the *Powertype* pattern.

5.1. Reference Ontology

During the process of reviewing the MiScOn ontology, the modeling of hierarchical structures in military organizations was refactored. The OntoUML extension proposed by Fonseca et al. [23] incorporated the MLT concepts and was therefore used in the reviewing of MiScOn to allow the representation of higher-order types and, at the same time, preserve the identity and temporal persistence of the entities.

The initial modeling of the structure of military organizations (MO) is shown in Figure 6(a). Note that MOs are classified according to several military organizational echelons, each with its own characteristics and relationships. For example, the 30^{th} Battalion, the 32^{nd} Battalion, and the 34^{th} Battalion are instances of the subtype *Battalion*, and are all commanded by the 11^{th} Brigade. Additionally, Battalions are commanded[4] by Brigades, Companies to Battalions, and so on. Moreover, each military organization has its own force (*strength* attribute), which represents the number of combatants it has.

The representation of organizational echelons in the military domain can benefit from the use of multi-level modeling seen in Sections 4.1 to 4.2. Promoting MO subtypes to *powertype* instances enables the definition of rules between these subtypes, which can be used to constrain MO instances. Additionally, it avoids the need for distinct relationships between each pair of the subtypes (subkinds), as presented by the authors in [24]. As seen in Figure 6(b), the model expresses the *isCommandedBy* relation, both at the *powertype* and at the *basetype*. In short, using the *Powertype* pattern simplified the model in Figure 6(a). Moreover, type-level attributes represented at the MO type were promoted to the *powertype* level as restrictions over the MO instances.

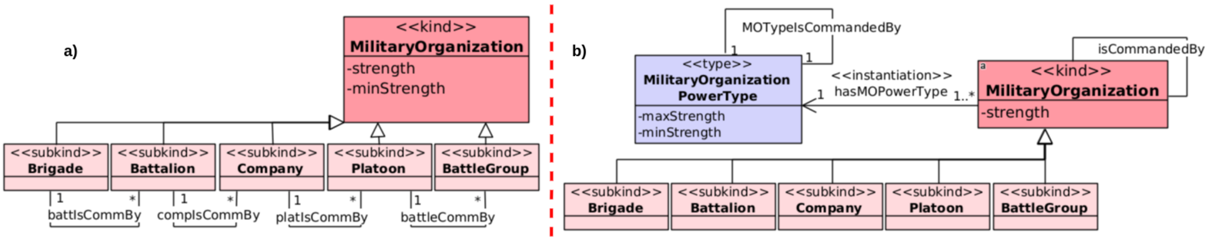


Figure 6: Fragment of the MiScOn reference ontology representing Military Organizations.

[4] It is important to highlight that the subordination relationship (isCommandedBy) in Figure 6 differs from the subordination concept of MLT (A is subordinate to B iff the instances of A are subtypes of instances of B), because it is related to the subordination relation in the military hierarchy.

Using the *powertype* `MilitaryOrganizationPowertype` (MOPT) and the basetype `MilitaryOrganization` (MO), it becomes possible to answer some CQs. CQ2: MOPT is the only type that partitions MO, through the *hasMOPowerType* relation. The cardinality (1 - 1..*) guarantees that for each MO instance, there is an instance of exactly one instance of MOPT. Conversely, for each MOPT instance, there is at least one instance of MO. CQ4 can also be answered, since MO can be partitioned according to MOPT instances ("Brigade", "Battalion", etc.). Moreover, restrictions on relationships and attributes are made explicit at MOPT, answering CQ5 and CQ6, respectively. According to MOPT, a *Battalion* can be commanded by exactly one *Brigade*. For example, given the MO 36^{th} *Battalion*, which is an instance of type (*hasMOPowerType*) *Battalion*, it can be commanded by the 11^{th} *Brigade*, and its *strength* attribute can assume values between *maxStrength* and *minStrength* values of the MOPT *Battalion* instance.

CQ2, CQ4 and CQ6 can also be answered in the modeling fragment on military vehicles and their various types, each of which can have different attributes depending on their respective *powertypes*, as shown in Figure 7. The *Kind* `Vehicle` is specialized in `Armored Vehicle`, which has, in turn, various subtypes, such as `Guarani`, `Urutu` and `Cascavel`. With the *VehiclePowerType*, it was possible to represent the regularity attributes at the power type level, restricting the instances of the *Armored Vehicle* type concerning the minimum and maximum passenger capacity and the maximum speed on land, according to the corresponding instance type.

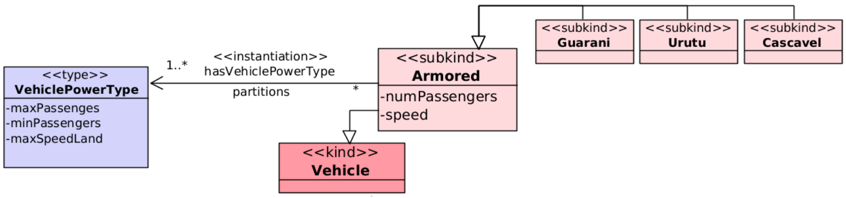


Figure 7: Fragment of the MiScOn reference ontology representing armored vehicles.

5.2. Operational Ontology

Following the development of the reference ontology in OntoUML, we implemented an operational ontology based on OWL language, incorporating instantiations and inference rules expressed in SWRL. To support multi-level modeling, we adopted the **punning** technique, which allows the same URI to be used as both a class and an individual. This enables reasoning across different levels of abstraction and supports rules that depend on treating entities as both types and instances. Furthermore, the lightweight version of UFO named gUFO has support to work with MLT.

```
<owl:Class rdf:about="https://example.com/miscon#Battalion">
        <rdfs:subClassOf rdf:resource="https://example.com/miscon#MilitaryOrganization"/>
</owl:Class>
<owl:NamedIndividual rdf:about="https://example.com/miscon#Battalion">
        <rdf:type rdf:resource="https://example.com/miscon/gufo#SubKind"/>
        <rdf:type rdf:resource="https://example.com/miscon/miscon#MilitaryOrganizationPowerType"/>
        <MOTypeIsCommandedBy rdf:resource="https://example.com/miscon/miscon#Brigade"/>
        <maxStrength rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">800</maxStrength>
        <minStrength rdf:datatype="http://www.w3.org/2001/XMLSchema#decimal">500</minStrength>
</owl:NamedIndividual>
```

Listing 1: Battalion being represented as a class and as an individual in the operational ontology in RDF/XML syntax.

This approach is aligned with Lenzerini's work [25], which explores meta-level queries in OWL2 QL using punning to enable querying both the structure and the instance level of an ontology. In our case, punning facilitated the operationalization of multi-level constructs, as illustrated in Listing 1, where the class `Battalion` is represented both as a class and an individual. The complete reference and operational ontologies are available in the project site[5].

The use of punning technique along with SWRL rules help on creating instantiation restrictions in the operational ontology. Listing 2 shows a rule in SWRL which enforces the fact that for one instance of a military organization (o_1) to be commanded by another (o_2), its corresponding *powertypes* must also be commanded by each other. It can be seen in Figure 6(b) that the relationship `isCommandedBy` is at the *basetype* level, while the relationship `MOTypeIsCommandedBy` is at the *powertype* level. Once the *powertype* level relationships are instantiated, then the relationship instantiations at the *basetype* level will be constrained by the specified rule.

```
MilitaryOrganization(?o2) ^ MilitaryOrganization(?o1) ^
isCommandedBy(?o2, ?o1) ^
differentFrom(?o1, ?o2) ^
MilitaryOrganizationPowerType(?ompt1) ^ MilitaryOrganizationPowerType(?ompt2) ^
hasMOPowerType(?o1, ?ompt1) ^ hasMOPowerType(?o2, ?ompt2) ^
-> MOTypeIsCommandedBy(?ompt2, ?ompt1)
```

Listing 2: SWRL rule to express the restriction on the isCommandedBy relationship between MOs, based on their powertypes.

6. Final Considerations and Future Works

This work proposed a novel approach for identifying characteristics in conceptual models that indicate the need for multi-level modeling. Focusing on the Powertype pattern within the context of Multi-Level Theory (MLT), we introduced a step-by-step process to make implicit modeling elements—such as attributes and relationships—explicit. Based on this process, we developed a set of domain-independent competency questions that reinforce the benefits of identifying and applying the Powertype pattern in various modeling scenarios. The proposed set is not exhaustive and their development remains an open area of research.

Our findings demonstrate that the Powertype pattern, when applied in light of MLT, significantly enhances semantic modeling. Furthermore, the proposed guidelines shows promise for supporting the process of ontological unpacking in existing conceptual models, where the identification of patterns and anti-patterns and the knowledge about how and when apply them is essential for achieving well-founded representations aligned with foundational ontologies.

Acknowledgments

This research has been supported by CAPES ("Coordenação de Aperfeiçoamento de Pessoal de Nível Superior"), and funded by FINEP/DCT/FAPEB (ref.: 2904/20 contract No. 01.20.0272.00) under the “System of Systems of Command and Control” project.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and DeepL Translate for grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] C. M. Fonseca, J. P. A. Almeida, G. Guizzardi, V. A. Carvalho, Multi-level conceptual modeling: Theory, language and application, Data & Knowledge Engineering 134 (2021) 101894.

[2] J. Odell, Power types, J. Object Oriented Program. 7 (1994) 8–12.

[5] https://github.com/comp-ime-eb-br/S2C2-IME

[3] L. Cardelli, Structural subtyping and the notion of power type, in: Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, 1988, pp. 70–79.

[4] G. Guizzardi, A. Botti Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. Prince Sales, Ufo: Unified foundational ontology, Applied Ontology 17 (2022) 167–210. doi:10.3233/AO-210256.

[5] H. Herre, General formal ontology (gfo): A foundational ontology for conceptual modelling, in: Theory and applications of ontology: computer applications, Springer, 2010, pp. 297–345.

[6] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, et al., Dolce: a descriptive ontology for linguistic and cognitive engineering (2003). URL: https://hdl.handle.net/20.500.14243/196545.

[7] N. Guarino, Bfo and dolce: So far, so close..., COSMOS + TAXIS 4 (2017).

[8] G. Guizzardi, Ontological patterns, anti-patterns and pattern languages for next-generation conceptual modeling, in: International Conference on Conceptual Modeling, Springer, 2014, pp. 13–27.

[9] G. Guizzardi, J. Almeida, N. Guarino, V. A. Carvalho, Towards an ontological analysis of powertypes, in: Proceedings of the Joint Ontology Workshops 2015 Episode 1: The Argentine Winter of Ontology co-located with the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015); Buenos Aires, Argentina, July 25-27, 2015, volume 1517, RWTH, 2015.

[10] N. Guarino, T. P. Sales, G. Guizzardi, Reification and truthmaking patterns, in: Conceptual Modeling: 37th International Conference, ER 2018, Xi'an, China, October 22–25, 2018, Proceedings 37, Springer, 2018, pp. 151–165.

[11] V. A. Carvalho, J. P. A. Almeida, Toward a well-founded theory for multi-level conceptual modeling, Software & Systems Modeling 17 (2018) 205–231.

[12] V. A. Carvalho, J. P. A. Almeida, C. M. Fonseca, G. Guizzardi, Extending the foundations of ontology-based conceptual modeling with a multi-level theory, in: Conceptual Modeling: 34th International Conference, ER 2015, Stockholm, Sweden, October 19-22, 2015, Proceedings 34, Springer, 2015, pp. 119–133.

[13] V. A. Carvalho, J. P. A. Almeida, G. Guizzardi, Using a well-founded multi-level theory to support the analysis and representation of the powertype pattern in conceptual modeling, in: Advanced Information Systems Engineering: 28th International Conference, CAiSE 2016, Ljubljana, Slovenia, June 13-17, 2016. Proceedings 28, Springer, 2016, pp. 309–324.

[14] J. P. A. Almeida, V. A. Carvalho, F. Brasileiro, C. M. Fonseca, G. Guizzardi, Multi-level conceptual modeling: Theory and applications, in: Proceedings of the XI Seminar on Ontology Research in Brazil and II Doctoral and Masters Consortium on Ontologies, October 1st-3rd, 2018., volume 2228, CEUR-WS, São Paulo, Brazil, 2018, pp. 26–41.

[15] B. Neumayr, K. Grün, M. Schrefl, Multi-level domain modeling with m-objects and m-relationships, in: Proceedings of the Sixth Asia-Pacific Conference on Conceptual Modeling-Volume 96, Citeseer, 2009, pp. 107–116.

[16] G. Guizzardi, J. P. A. Almeida, Stability patterns in ontology-driven conceptual modeling, in: Proceedings of the XIII Seminar on Ontology Research in Brazil and IV Doctoral and Masters Consortium on Ontologies (ONTOBRAS 2020), volume 2728, CEUR-WS, 2020, pp. 148–160.

[17] J. D. Lara, E. Guerra, J. S. Cuadrado, When and how to use multilevel modelling, ACM Transactions on Software Engineering and Methodology (TOSEM) 24 (2014) 1–46.

[18] T. Halpin, Subtyping revisited, CEUR Workshop Proceedings 365 (2012).

[19] G. Guizzardi, N. Guarino, Explanation, semantics, and ontology, Data Knowledge Engineering 153 (2024) 102325. URL: https://www.sciencedirect.com/science/article/pii/S0169023X24000491. doi:https://doi.org/10.1016/j.datak.2024.102325.

[20] A. M. Demori, An Ontology-Based Approach for Reproducing Military Operation Scenarios [Uma Abordagem Baseada em Ontologia para Reprodução de Cenários de Operações Militares], Master's dissertation, Military Institute of Engineering, Rio de Janeiro, Brazil, 2023.

[21] R. Falbo, Sabio: Systematic approach for building ontologies, in: Proceedings of the 1st Joint Workshop ONTO.COM / ODISE on Ontologies in Conceptual Modeling and Information Systems Engineering co-located with 8th International Conference on Formal Ontology in Information Systems (FOIS 2014), volume 1301, CEUR Workshop Proceedings, Rio de Janeiro, Brazil, 2014.

[22] G. Guizzardi, Ontological foundations for structural conceptual models, Ph.D. thesis, University of Twente, 2005.

[23] C. M. Fonseca, G. Guizzardi, J. P. A. Almeida, T. P. Sales, D. Porello, Incorporating types of types in ontology-driven conceptual modeling, in: International Conference on Conceptual Modeling, Springer, 2022, pp. 18–34.

[24] A. M. Demori, J. C. C. Tesolin, M. C. R. Cavalcanti, D. F. C. Moura, Supporting simulation of military communication systems using well-founded modeling., in: ONTOBRAS, 2022, pp. 73–84.

[25] M. Lenzerini, L. Lepore, A. Poggi, Metamodeling and metaquerying in owl 2 ql, Artificial Intelligence 292 (2021) 103432.

Integração de dados de saúde pública para apoio à tomada de decisão: uma abordagem baseada em grafo de conhecimento semântico

Sandro de Carvalho Franco[1], Nacles Bernardino Pirajá Gomes[1] and
Laís do Nascimento Salvador[1]

[1]*Programa de Pós-graduação em Ciências da Computação (PGCOMP)*
Universidade Federal da Bahia - Salvador/BA

Resumo

Este artigo aborda a complexidade e os desafios dos sistemas de informação em saúde pública no Brasil, caracterizados por bases de dados diversas e desconectadas. Para solucionar o problema do acesso unificado a dados distribuídos em múltiplas plataformas, propomos uma abordagem inovadora baseada no uso de ontologias. Ao definir um vocabulário comum que estabelece correspondências semânticas entre diferentes bases de dados, as consultas tornam-se mais simples. O texto descreve a implementação de uma integração semântica apoiada por Grafos de Conhecimento Semântico (Semantic Knowledge Graphs — SKGs). O estudo concentra-se em uma ontologia desenvolvida para responder às principais questões levantadas por especialistas em saúde pública. O trabalho define indicadores baseados em dados críticos dos principais sistemas do Sistema Único de Saúde (SUS), incluindo informações sobre nascimentos, óbitos e notificações de COVID-19, utilizando o município de Camaçari, Bahia, como estudo de caso. O objetivo é permitir que gestores de saúde analisem indicadores essenciais, como mortalidade, imunização, hospitalizações e notificações, destacando a eficácia dos SKGs na melhoria da gestão em saúde pública. Dessa forma, integramos fontes de dados relacionais em uma base unificada e virtualizada, onde consultas podem ser executadas para responder às diferentes Questões de Competência (QCs) propostas pela equipe técnica de Camaçari-BA.

Keywords

Integração de dados, grafos de conhecimento semântico, ontologias, bases de dados heterogêneas.

Abstract

This paper addresses the complexity and challenges of public health information systems in Brazil, which are characterized by diverse and disconnected databases. To solve the problem of unified access to data spread across multiple platforms, we propose an innovative approach using ontologies. By defining a common vocabulary that establishes semantic correspondences among different databases, queries become simpler. The text describes the implementation of semantic integration supported by Semantic Knowledge Graphs (SKGs). The study focuses on an ontology developed to respond to key questions raised by public health specialists. The project defines indicators based on critical data from the main systems of the Unified Health System (Sistema Único de Saúde - SUS), including information on births, deaths, and COVID-19 notifications, using data from the municipality of Camaçari, Bahia, as a case study. The goal is to allow health administrators to analyze essential indicators such as mortality, immunization, hospitalizations, and notifications, highlighting the effectiveness of SKGs in improving public health management. In this way, we integrate relational data sources into a unified and virtualized database, where queries can be executed to answer different Competency Questions (CQs) proposed by the technical team of Camaçari-BA.

Keywords

Data integration, semantic knowledge graphs, ontologies, heterogeneous databases.

1. Introdução

Nas últimas décadas, o conceito de ontologia vem adquirindo destaque em diversas áreas do conhecimento, transpondo suas raízes filosóficas para se tornar uma ferramenta essencial na ciência da computação. No contexto da organização e interpretação de dados, as ontologias fornecem uma estrutura formal e compartilhada para representar o conhecimento de um domínio, definindo um conjunto

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

✉ sandrofranco@ufba.br (S. d. C. Franco); nacles.gomes@ufba.br (N. B. P. Gomes); laisns@ufba.br (L. d. N. Salvador)
iD 0009-0003-3497-1171 (S. d. C. Franco); 0000-0002-4326-2188 (N. B. P. Gomes); 0000-0001-7441-057X (L. d. N. Salvador)

estruturado de termos e suas relações (Guarino et al. (2009), Grenon and Smith (2011)). Essa capacidade de integrar semanticamente informações de fontes, domínios e contextos diversos tornou-se criticamente relevante em um cenário de crescimento exponencial de dados, permitindo não apenas a organização conceptual, mas também a inferência e a extração de conhecimento novo, facilitando a interoperabilidade entre sistemas de informação heterogêneos (Grenon and Smith, 2011, de Cerqueira, 2016).

Este estudo insere-se em um contexto de inovação, propondo uma metodologia avançada para a integração de dados heterogêneos, com foco especial no Sistema Único de Saúde (SUS) dos municípios brasileiros. Por meio da construção de um Grafo de Conhecimento Semântico (GCS), este trabalho visa superar os desafios impostos pela diversidade das fontes de dados. Tal abordagem promove uma compreensão mais aprofundada e acessível das informações relativas a registros de óbitos, nascimentos e notificações de casos de COVID-19. Expandindo a pesquisa anterior realizada por (Gomes et al., 2022), este estudo aplica uma metodologia híbrida onde um vocabulário compartilhado acessa a camada ontológica, que por sua vez, acessa as fontes de dados, conforme caracterizado por (Ekaputra et al., 2017), para a elaboração do GCS. Esta técnica destaca o valor dessas ferramentas em descobrir conexões e conhecimentos implícitos nos dados.

Portanto, este estudo objetiva desenvolver um Grafo de Conhecimento Semântico (GCS), ONTOVID II, para integrar bases de dados heterogêneas do Sistema Único de Saúde (SUS) — especificamente do SIM (Sistema de Informações de Mortalidade), SINASC (Sistema de Informações de Nascidos Vivos) e e-SUS Notifica (Sistema de Notificação do Ministério da Saúde). A proposta visa fornecer aos gestores municipais uma visão unificada, servindo como uma ferramenta crucial para a tomada de decisão em saúde pública. Consequentemente, a questão de pesquisa que norteia esta investigação é: Como a integração semântica via GCS pode transformar dados dispersos das bases como SIM, SINASC e e-SUS Notifica em conhecimento acionável para a gestão municipal de saúde?

A próxima seção apresenta abordagens para integração de dados utilizando ontologias, conceitos sobre *OBDA (Ontology-Based Data Access)*, *OBDI (Ontology-Based Data Integration)* e Grafos de Conhecimento Semântico (GCS). A seção 3 traz a revisão literária de trabalhos relacionados ao tema proposto. Já a seção 4 apresenta as fontes de dados utilizadas e a construção das ontologias. A seção 5 aborda a definição e utilização do GCS. A seção 6 apresenta os resultados obtidos e a seção 7 traz a conclusão do texto e proposta de trabalhos futuros.

2. Integração de Dados

Considerado um dos problemas mais antigos relacionado a abordagens com dados, o estudo sobre integração teve seu início na década de 90, desde então, uma grande variedade de aspectos referentes a estes problemas foram pesquisados tanto na área acadêmica quanto na indústria, dentre eles a possibilidade de utilizar a integração semântica de dados. Usar semântica significa desenvolver sistemas de integração de dados onde a semântica dos dados é explicitamente especificada e é levada em consideração para o planejamento de todas as funcionalidades do sistema (Flesca et al., 2018). Nas últimas duas décadas, essa ideia se tornou cada vez mais crucial para uma ampla variedade de aplicativos de processamento de informações e tem recebido muita atenção nas comunidades de IA, banco de dados, web e mineração de dados (Natalya F. Noy, 2005). Nesse aspecto, o uso de ontologias, como artefato computacional que modela estruturas de sistemas de informação, vem sendo aplicado na web semântica, promovendo a interoperabilidade entre sistemas. Ontologias computacionais, formalizadas e compartilhadas, são essenciais para organizar e gerenciar conhecimento, definindo entidades e suas relações de maneira que possam ser processadas por máquinas. Essas estruturas não apenas facilitam a comunicação e compreensão entre sistemas e usuários, mas também auxiliam em todas as etapas do ciclo de vida do software, desde a análise de requisitos até a manutenção, por meio de uma especificação clara e consensual dos conceitos e suas inter-relações (Calhau and Falbo, 2010). A consulta a dados utilizando ontologias podem ser realizadas em bases que lidam com o mesmo domínio, neste caso a consulta pode ser mais fácil de ser realizada pois o contexto seria o mesmo, ou sobre bases que lidam com domínios diferentes, o que poderia tornar as consultas mais difíceis de se realizar. Para os casos mais complicados, pode-se utilizar integração semântica sobre os dados a serem consultados com o objetivo de assegurar que informações distintas, armazenadas em bases distintas, possam ser integradas por possuírem equivalência semântica, ou seja, são relacionados ao mesmo assunto.

Segundo (Gardner, 2005), a chave para ser capaz de integrar informações de forma reutilizável é o uso da semântica, que descreve o significado de uma palavra ou conceito. Neste projeto foi possível responder questões referentes às bases SIM, SINASC e e-SUS Notifica, tais como, por exemplo: Quantos foram os indivíduos imunizados que vieram a óbito por COVID-19 (SIM x COVID-19), Quantos foram os indivíduos nascidos e contaminados com COVID-19 (SINASC x COVID-19), Quantos foram os indivíduos nascidos e contaminados com COVID-19 que vieram a óbito (SIM x SINASC x COVID-19), entre outras possibilidades. Embora a palavra indivíduos seja utilizada em diversas questões, é possível perceber que, mesmo entre bases distintas, há um conceito comum compartilhado entre elas. A partir desse conceito, torna-se viável inferir informações relevantes a respeito desses dados heterogêneos. (Gardner, 2005) diz que a capacidade de distinguir entre estes sinônimos, homônimos e termos relacionados é essencial ao integrar dados de diferentes repositórios.

2.1. OBDA (Ontology-Based Data Access)

Organizações complexas geralmente possuem grande quantidade de dados armazenados em diferentes bases, o que demanda a implementação de consultas unificadas. É o que acontece com as secretarias municipais e estaduais de saúde no acesso às bases do SUS. Uma forma de mitigar este problema é usar a abordagem OBDA que conecta uma ontologia (TBox) a dados de um banco de dados relacional (ABox) por meio de uma camada de mapeamento, que posteriormente pode ser utilizada por um raciocinador automatizado, como ilustrado na Figura 1. Segundo (Bagosi et al., 2014) *ODBA* é um paradigma de acesso a dados por meio de uma camada conceitual. Essa camada é expressa na forma de uma ontologia, no formato *Web Ontology Language* (OWL), e os dados armazenados em bancos de dados relacionais.

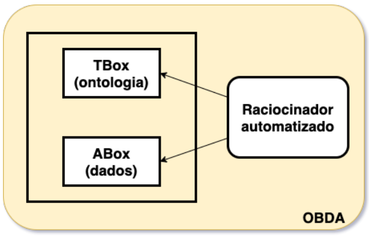


Figura 1: *Esquema OBDA* (Keet, 2020)

2.2. OBDI (Ontology-Based Data Integration)

Considerando um exemplo em que mesma pessoa é representada em um banco de dados por meio de nome e data de nascimento, num segundo banco de dados por meio do CPF (Cadastro de Pessoa Física) e num terceiro ela seja representada pela CNH (Carteira Nacional de Habilitação). Com a intenção de combinar informações sobre esta mesma pessoa que estão armazenados em bancos de dados diferentes, é preciso encontrar uma maneira em que se possa representá-la e combinar suas informações. Para este tipo de integração de dados armazenados em bancos de dados distintos pode-se utilizar a abordagem *Ontology-Based Data Integration* (OBDI). Segundo (Calvanese et al., 2015), OBDI é uma extensão do OBDA em que os dados não são armazenados em um único banco de dados, mas em uma infinidade de bancos de dados que precisam ser consultados de forma integrada, mas mantendo a mesma arquitetura conceitual baseada em mapeamentos. Uma das principais características dos modelos que empregam o OBDI é a independência conceitual em relação aos dados, o que permite agregar novas visualizações e fontes de dados. O modelo OBDI possui quatro abordagens que podem auxiliar na construção de uma ontologia. Para (Wache et al., 2002) as principais abordagens podem ser classificadas como: a) abordagem de ontologia única; b) abordagem de múltiplas ontologias; e c) abordagem híbrida. Existe uma quarta abordagem chamada de GAV (Global-as-View), variante adicional proposta por (Ekaputra et al., 2017).

2.3. Grafos de Conhecimento Semântico - GCS

(Soylu et al., 2018) nos definem que os GCS são um mecanismo para consolidar e integrar semanticamente um grande número de fontes de dados heterogêneas em um espaço de dados abrangente. Os GCS são reconhecidos como uma abordagem eficaz no processo de integração semântica de dados heterogêneos, apesar de serem considerados como um paradigma relativamente novo. As ontologias desempenham um papel essencial neste cenário sendo utilizadas como um vocabulário unificado que permite a combinação e enriquecimento de informações para a realização de consultas complexas e inferências sobre os dados integrados (Calvanese et al., 2018). Segundo (Junior, 2021) um GCS é destinado a acumular e transmitir conhecimento do mundo real, cujos nós representam entidades de interesse e cujas arestas representam relações entre essas entidades. Dentre as múltiplas tarefas envolvidas no processo de integração de dados, uma das mais importantes é a compreensão dos conceitos e relações que existem por trás dos registros (Cheatham and Pesquita, 2017).

3. Trabalhos Relacionados

Para iniciar o desenvolvimento deste trabalho foi feita uma revisão na literatura buscando trabalhos que tivessem relação direta ou indireta com o tema da pesquisa. Inicialmente, examinamos pesquisas que demonstravam a aplicação da integração semântica em domínios variados. Por exemplo, (Auceli et al., 2019) exploraram a correlação entre rendimento escolar e ocorrências policiais em Curitiba, integrando dados de educação e segurança pública. No domínio jurídico, (de Oliveira, 2017) aplicou ontologias para recuperar acórdãos do Supremo Tribunal Federal (STF), oferecendo uma alternativa aos formulários tradicionais. Em um escopo mais técnico e generalista, trabalhos como os de (Hajmoosaei and Abdul-Kareem, 2007), (Zhao and Ichise, 2014) e (Xiao et al., 2018) propuseram frameworks e soluções para integração de múltiplos bancos de dados relacionais, abordando desafios de heterogeneidade semântica e de esquema.

No domínio público não específico de saúde, por exemplo, o estudo de (Vidal et al., 2021) se destaca por propor a construção de um Grafo de Conhecimento Semântico (GCS) a partir de fontes heterogêneas como Receita Federal, IBGE e Correios, demonstrando a viabilidade técnica para criar uma representação semântica unificada de dados governamentais. Dentre os trabalhos focados em saúde pública, identificamos importantes esforços que servem como base para esta pesquisa. (Pereira, 2019) e a plataforma SemanticSUS [1] (Lima da Cruz et al., 2019) se concentram na integração semântica entre os sistemas SIM e SINASC, estabelecendo uma base metodológica crucial para a unificação de dados vitais. Aproximando-se do contexto pandêmico, (Maddalena and Baião, 2021) desenvolveram uma ontologia para a COVID-19 baseada na UFO (Unified Foundation Ontology), oferecendo um modelo conceptual valioso. Complementarmente, (Fernantes, 2012) visou à integração de dados epidemiológicos para minimizar a intervenção humana na análise, objetivo que dialoga com a motivação de automatização deste projeto.

Embora essa base literária seja robusta, observa-se uma lacuna crítica: a integração simultânea e tríplice entre dados vitais (SIM, SINASC) e dados epidemiológicos de notificação (e-SUS Notifica). Os trabalhos relacionados concentram-se primordialmente na integração de pares de bases ou na modelagem de um domínio único. Portanto, este projeto diferencia-se ao propor um GCS que consolida e inter-relaciona essas três fontes heterogêneas, com o propósito explícito de gerar insights acionáveis para a tomada de decisão municipal em tempo crítico, bem como facilitar a inclusão de novas fontes de dados do SUS, ampliando o GCS.

[1] https://semanticsus.github.io/semanticSUS/index.html

4. Fontes de dados

Para dar início à implementação das ontologias inicialmente firmou-se uma parceria entre a UFBA e a Secretaria Municipal de Saúde de Camaçari. Esta parceria foi necessária para que os dados pudessem ser acessados e com isso a ontologia pudesse ser construída com o objetivo de ser utilizada na Secretaria. Seguindo a Lei Geral de Proteção de Dados (LGPD), Lei nº 13.709/2018, este trabalho não apresenta dados pessoais, apenas informações quantitativas. As bases utilizadas neste trabalho foram as do SIM, SINASC e e-SUS Notifica, elas representam a Camada de Fonte de Dados da Figura 2. Na figura pode-se observar o modelo proposto para este projeto onde, em verde, tem-se a inclusão da ontologia de Notificações de COVID-19 e em azul a realização da integração semântica entre as ontologias SINASC x SIM x COVID-19. Com estas integrações, na Camada de Ontologia de Domínio (OD) encontra-se a ontologia capaz de responder questões de competência referentes às fontes de dados apresentadas. As base do SIM e SINASC foram configuradas em um banco de dados Firebird[2], rodando no sistema operacional Windows[3]. Por se tratar de um banco de dados antigo, foram disponibilizados dois arquivos de extensão **.fdb**, um para cada base de dados, não foi disponibilizado o acesso direto a base da secretaria. Para acesso a base de dados do e-SUS Notifica, foi disponibilizada uma *view* em PostgreSQL[4] que roda num servidor Linux OpenSUSE[5].

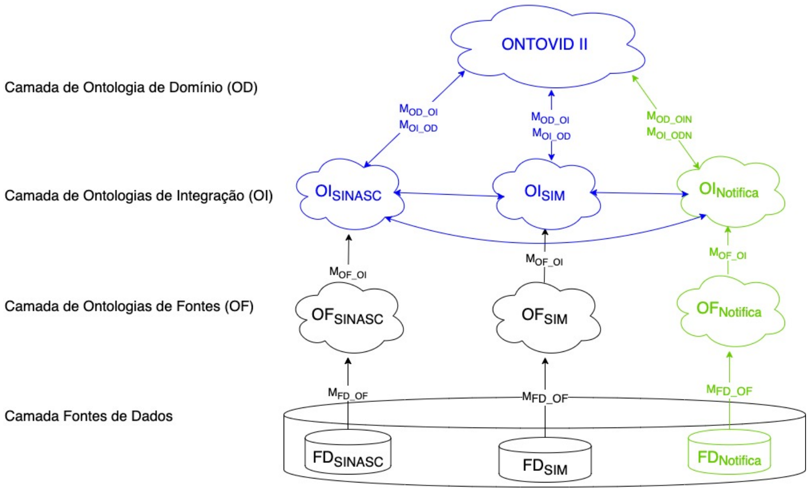


Figura 2: Fluxo proposto para este projeto. (Fonte: autor)

4.1. SINASC

A base de dados do SINASC contém informações sobre nascimentos e, assim como os dados relacionados ao SIM, apenas a tabela principal **TB_DN** (tabela sobre dados de nascimento) foi utilizada. Desta tabela consegue-se obter informações necessárias para construir a ontologia de fonte OF_SINASC, são elas: APGAR[6] primeiro minuto, APGAR quinto minuto, CNS, data de nascimento, ID, idade, nome completo, número de declaração de nascido vivo, peso ao nascer, quantidade de nascidos mortos, sexo, tempo de gestação e tipo de parto.

[2] https://firebirdsql.org/

[3] https://www.microsoft.com/pt-br/windows

[4] https://www.postgresql.org/

[5] https://www.opensuse.org/

[6] APGAR (Aparência, Pulso, Gesticulação, Atividade, Respiração) é uma escala que avalia a saúde do recém-nascido nos primeiros minutos de vida

4.2. SIM

A base de dados do SIM contém informações sobre mortalidade, para este projeto não foi utilizada toda a estrutura de dados do SIM, apenas a tabela principal **TB_DO**. Desta tabela consegue-se obter informações necessárias para construir a ontologia de fonte OF_SIM, são elas: CID, CNS, data do atestado de óbito, data do óbito, data de nascimento, dias vividos, ID, idade, nome completo, número de declaração de nascido vivo, número da declaração do óbito, quantidade de nascidos mortos e sexo.

4.3. e-SUS Notifica

A base de dados do e-SUS Notifica contém as informações de notificação de COVID19. Esta base trata-se de uma *view* que contém informações necessárias para construção da OF_NOTIFICA, são elas: bairro, CEP, cidade, classificação final, CNS, ID, CPF, data do teste, data do teste Lamp, data do teste PCR, data de nascimento, data da primeira dose, data da segunda dose, CBO, e-mail, estado, estado do teste, evolução do caso, laboratório da primeira dose, laboratório da segunda dose, nome completo, número da notificação, primeira dose, segunda dose, resultado do teste, sexo, sintoma e tipo do teste.

4.4. Ontologias de Fonte

Neste trabalho, as ontologias de fonte são as ontologias que representam as fontes de dados a elas associadas, sendo assim tem-se OF_SIM, OF_SINASC e OF_NOTIFICACOES, estando elas relacionadas as bases dados do SIM, SINASC e e-SUS Notifica respectivamente. Todas as ontologias foram criadas utilizando a linguagem OWL 2 com sintaxe RDF/XML e estas representam a Camada de Ontologias de Fonte(OF). A da Figura 3 mostra um exemplo de ontologia de fonte, a OF_SIM. Nesta ontologia temos as classes: i) **CausaDaMorte**, que representa a causa do óbito; ii) **Óbito**, que possui as informações sobre o óbito de uma pessoa; iii) **Pessoa**, que representa uma pessoa na base do SIM; iv) **Mãe**, que possui dados sobre a mãe de uma pessoa que veio a óbito e; v) **Pai**, que possui dados sobre o pai de uma pessoa que veio a óbito. A classe **Pessoa** tem as subclasses (**has subclass**) **Mae** e **Pai**, as linhas pontilhadas representam as relações entre as classes, onde i) Pessoa **temRegistroDeObito** da classe Obito; ii) Pessoa pode ter informações do pai (**temPaiSIM**); iii) Pessoa tem informações da mãe (**temMaeSIM**) e; iv) Obito **temCausaMorte** da classe CausaMorte.

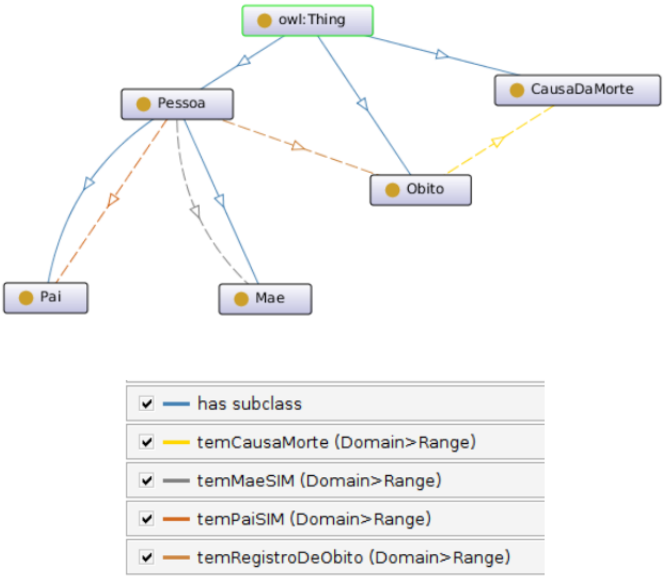


Figura 3: Representação da ontologia de fonte OF_SIM construída no Protégé. (Fonte: autor)

4.5. Ontologias de Integração

As ontologias de integração criadas para este projeto dão continuidade ao trabalho iniciado com as ontologias de fonte. As ontologias de integração também foram criadas no Protégé onde, cada ontologia foi criada a partir da importação da ontologia de fonte correspondente. Assim temos: i) OI_SIM importada de OF_SIM; ii) OI_SINASC importada de OF_SINASC e iii) OI_NOTIFICACOES importada de OI_NOTIFICACOES. As ontologias de integração construídas representam a Camada de Ontologias de Integração (OI) da Figura 2 da Proposta. Utilizando recurso de importação, qualquer alteração que seja realizada na ontologia de fonte automaticamente se reflete na ontologia de integração, isso reduz a necessidade de alteração em ambas ontologias. Diferentemente das ontologias de fonte, as ontologias de integração possuem ligação com bases de dados onde o Prótégé se conecta às bases utilizando *drivers JDBC*. A estratégia de vinculação de registros utilizada na Ontovid II baseia-se no conceito de "indivíduos" como elementos de referência compartilhada entre as bases. Para cada entidade, são priorizados identificadores disponíveis como CPF, data de nascimento, sexo e município de residência. Nos casos em que algum identificador está ausente, a instância é criada apenas com os atributos disponíveis, preservando a consistência semântica, ainda que com menor capacidade de seleção. Não foram aplicados, nesta fase, algoritmos avançados de resolução de conflitos, aspecto reconhecido como limitação e direcionado para evolução em trabalhos futuros.

5. ONTOVID II

A necessidade de produzir informações integradas e de fácil acesso tem impulsionado o desenvolvimento de novas soluções que facilitem o processo de integração entre estes dados. Este tipo de integração pode ser complexa e onerosa devido à heterogeneidade semântica ou distribuição de dados, e isso também acontece com dados da área de saúde. Soluções para este tipo de integração tem se mostrado eficiente com abordagens que utilizam ontologias. Este trabalho tem como objetivo realizar a integração semântica entre as bases SIM, SINASC e e-SUS Notifica por meio de grafos de conhecimento semântico, de forma a minimizar as dificuldades recorrentes enfrentadas por gestores do SUS no acesso a dados de saúde integrados. A proposta visa disponibilizar indicadores consolidados que apoiem análises relacionadas a nascimentos, óbitos e notificações de COVID-19, fortalecendo a capacidade de planejamento e gestão em saúde pública. Para isso foram utilizadas soluções semelhantes para criação da ontologia com domínio relacionado a dados sobre Nascimentos, Óbitos e Notificações de COVID-19. Para além, utilizou-se também o cenário 4 da metodologia NeOn [7], neste processo foram desenvolvidas as ontologias determinando seu domínio e escopo. Uma vez construídas as ontologias de fonte e de integração, o passo subsequente foi a implementação da ontologia de domínio a ser utilizada pelo Grafo de Conhecimento Semântico (GCS). No modelo proposto, representado na Camada Semântica da Figura 6, esta ontologia de domínio corresponde à camada ilustrada na Figura 2. A ontologia de domínio resultante, denominada ONTOVID II e representada na Figura 4, consolida os conceitos das três fontes de dados (e-SUS Notifica, SINASC e SIM).

Sua estrutura pode ser analisada da seguinte forma: a) *Bloco 1:* Agrupa as classes e subclasses necessárias para representar o domínio do e-SUS Notifica, como TesteCovid19, Notificacao, ProfissionalDeSaude e Vacina; b) *Bloco 2:* Contém as classes e subclasses referentes aos dados do SINASC, como RegistroNascidoVivo, Gestacao, Nascimento e Mae; e c) *Bloco 3:* Apresenta as classes e subclasses do SIM, centradas na classe Obito e CausaDaMorte. Todas essas classes herdam propriedades e restrições de suas respectivas ontologias locais, garantindo a consistência semântica. Finalmente, as classes Pessoa, Pai e Mae (à direita) desempenham um papel crucial como classes de unificação, agregando e harmonizando as propriedades comuns provenientes das três ontologias locais, o que permite uma visão integrada e coerente do cidadão nos diferentes sistemas.

[7] Cenário 4 da metodologia NeOn é um dos nove cenários propostos para o desenvolvimento de ontologias, focado especificamente na reutilização e reengenharia de recursos ontológicos não ontológicos.

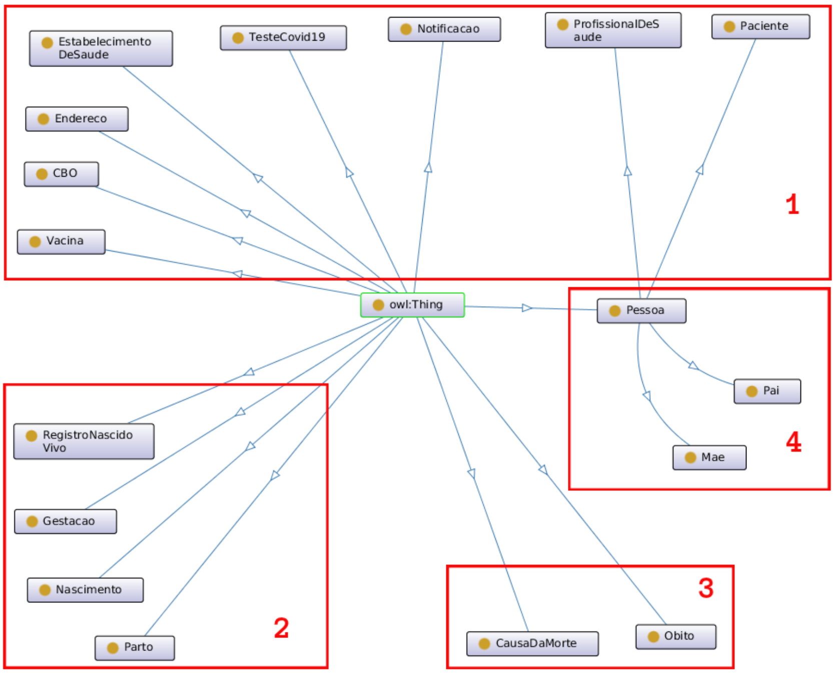


Figura 4: Ontologia de domínio Ontovid II. (Fonte: autor)

A Figura 5 mostra o modelo conceitual do GCS que representa a ontologia de domínio presente na Figura 2 e, a seguir serão apresentadas as classes das ontologias locais, que permearam o acesso as fontes de dados do SIM, SINASC e e-SUS Notifica:

- SIM: i) Óbito e; ii) Causa da Morte
- SINASC: i) Gestação; ii) Parto; iii) Nascimento e; iv) Registro de Nascido Vivo
- e-SUS Notifica: i) Paciente; ii) Endereço; iii) Profissional de Saúde; iv) Vacina; v) Notificação; vi) Classificação Brasileira de Ocupação (CBO) e vii) Estabelecimento de Saúde

Após levantamento dos modelos conceituais das bases, observou-se que as classes **Pessoa**, **Pai** e **Mãe** são conceitos que se relacionam a todas ontologias das fontes de integração. Com relação ao SIM observa-se que *Pessoa* que tem *Mãe*, tem *Pai*, tem registro de *Óbito* e este tem a *Causa da Morte*. Com relação ao SINASC, uma *Pessoa* nascida possui *Registro de Nascido Vivo* e neste registros encontram-sem informações do *Parto*, do *Nascimento*, da *Gestação*, da *Mãe* e do *Pai*. Com relação às Notificações, *Profissional de Saúde* é uma *Pessoa* que aplica *Vacina* e o mesmo pode se vacinar, tem *CBO*, realiza as *Notificações* e tem vínculo com *Estabelecimento de Saúde*. *Paciente* é uma *Pessoa* que tem dados da *Notificação* e a notificação tem o registro de *Vacina* e do *TesteCovid19*.

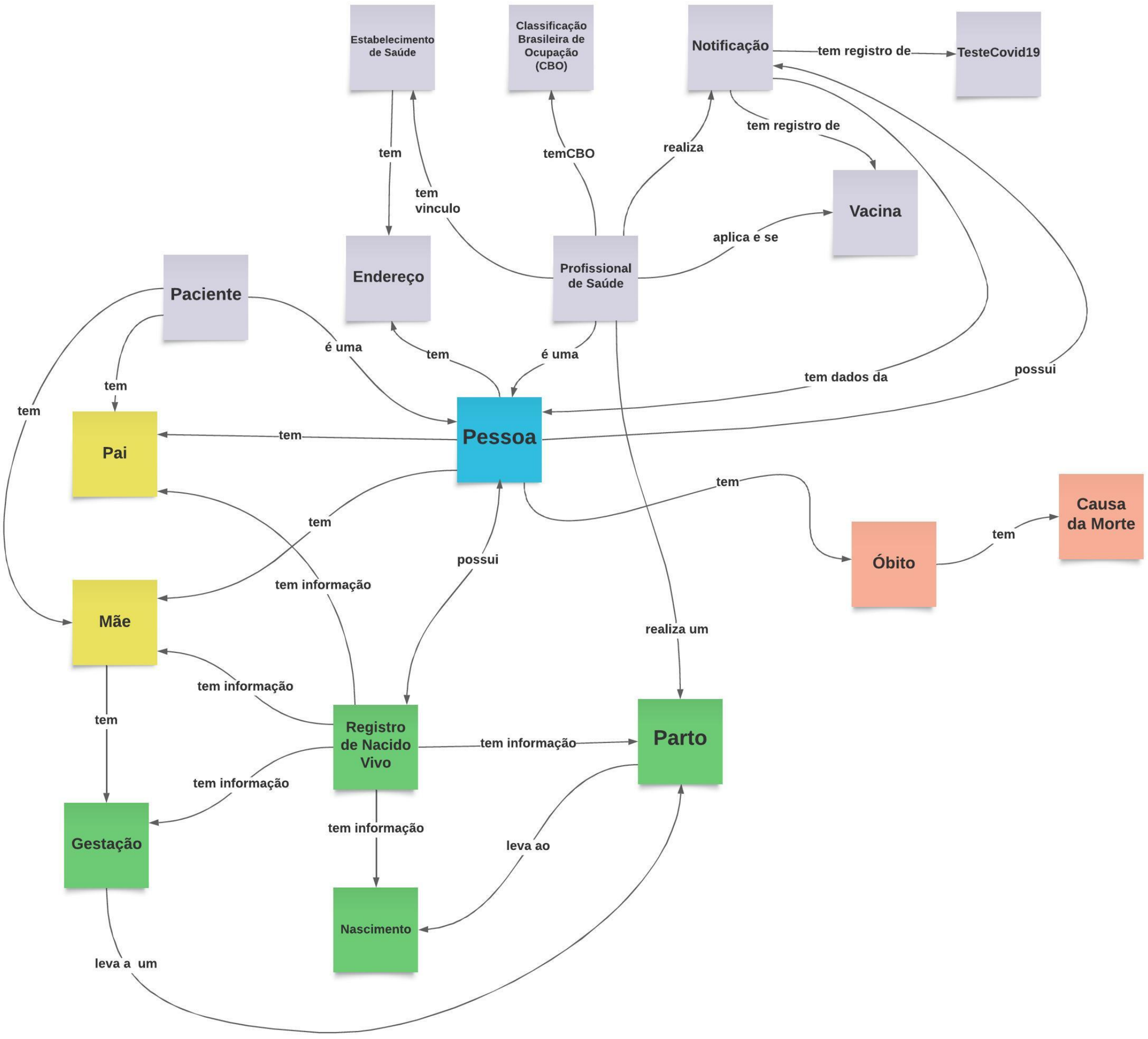


Figura 5: Modelo conceitual do GCS (Fonte: autor)

A Figura 6 a seguir mostra o modelo em camadas da proposta do projeto. Este modelo foi construído baseado na abordagem hibrida apresentada na seção 2.2. No modelo apresentado é possível observar a *Camada de Aplicações e Consultas* onde são realizadas consultas SPARQL, a *Camada Semântica* onde encontra-se o GCS Compartilhado que representam o modelo apresentado na Figura 5, a *Camada Ontológica* que representam as ontologias que construídas para o projeto e por fim a *Camada de Dados* que representa as bases acessadas pelas ontologias da camada ontológica.

6. Resultados da ONTOVID II

Para validar a integração realizada e consequentemente a ontologia de domínio, foram utilizadas Questões de Competência (QCs) para extrair informações das bases utilizadas (SIM, SINASC e Notificações de COVID-19). Para isso, tomou-se como base painéis que mostram graficamente dados de saúde utilizado pela Secretaria Municipal de Saúde de Camaçari, são eles: i) Painel de Mortalidade (SIM); ii) Painel de Natalidade (SINASC); iii) Painel de Notificações de COVID-19; iv) Painel de Saúde Materno-Infantil (SIM - SINASC) e; v) Painel de Notificações x SIM x SINASC. Para atender a estes painéis foram formuladas **quatorze** questões de competência para o Painel de Natalidade, **sete** para o Painel de Mortalidade e **seis** para o painel de Saúde Materno-Infantil, a Tabela 1 traz algumas dessas questões de competência. Foram formuladas também questões que abrangessem os dados de Notificação de COVID-19.

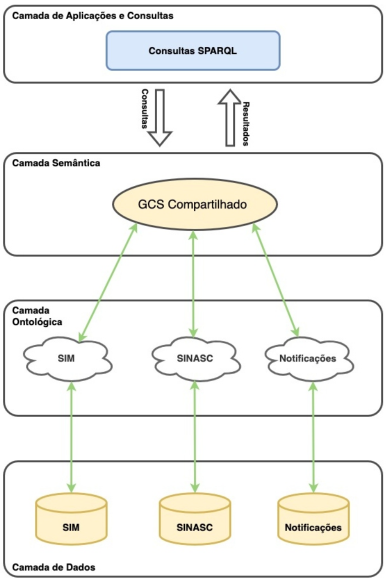


Figura 6: Modelo ONTOVID II em camadas.(Fonte: autor).

Para este caso foram formuladas **seis** Questões de Competência (QCs) envolvendo somente a base de Notificações e **oito** envolvendo as bases do SIM, SINASC e Notificações. Seguem as Questões de Competência levantadas para este projeto e seus resultados em 19 de julho de 2023. Observando os resultados relacionados na Tabela 1, vale ressaltar que as consultas realizadas para compor o Painel de Mortalidade (SIM), o Painel de Natalidade (SINASC) e o Painel de Notificações de COVID-19 podem ser consideradas consultas simples pois suas ontologias acessam diretamente uma única base de dados. Já as consultas realizadas para compor o Painel de Saúde Materno-Infantil (SIM - SINASC) e o Painel de Notificações x SIM x SINASC podem ser consideradas consultas complexas pois estas acessam duas ou mais bases de dados diferentes. Com relação a estas consultas mais complexas, o uso de ontologias se torna um facilitador pois seu dicionário de dados abstrai as relações entre estas bases de dados distintas, assim, conhecendo este dicionário, construir consultas que relacionem estas bases se torna uma tarefa mais simplificada de se realizar. As Questões de Competência (QCs) descritas na Tabela 1 foram construídas em linguagem SPARQL e executadas no Prótégé. Todos os resultados foram apresentados à Secretaria Municipal de Saúde de Camaçari para validação dos gestores. As consultas foram executadas no ambiente de desenvolvimento e os responsáveis pela validação analisaram e aceitaram[8] como válidos os dados informados nas consultas, com isso, há um forte indício para utilização da ONTOVID-II em ambiente de produção. A avaliação realizada concentrou-se na validação qualitativa dos resultados junto à equipe técnica da Secretaria Municipal de Saúde de Camaçari. Essa abordagem foi suficiente para demonstrar a correção das consultas formuladas, mas não contemplou métricas quantitativas como tempo médio de execução, escalabilidade ou indicadores de precisão no processo de vinculação de registros. O reconhecimento dessas limitações é importante para orientar pesquisas futuras, nas quais se pretende ampliar os experimentos e adotar medidas comparativas de desempenho.

[8] https://drive.google.com/drive/folders/1Ow_Y41NJERS2TWEGGK71n9U_zgVUzEH1?usp=sharing

Tabela 1
Questões de Competência para ONTOVID II
(Fonte: autor)

Painel de Mortalidade (SIM)	**Resultado**
QC 01: Número de óbitos nos anos 2021, 2022 e 2023	2021: 2088 2022: 1580 2023: 576
QC 02: Mortalidade por COVID-19 - CID B342	Total: 799
Painel de Natalidade (SINASC)	**Resultado**
QC 01: Número de nascidos vivos nos anos 2021, 2022 e 2023	2021: 3880 2022: 2220 2023: 1505
QC 02: Taxa bruta de natalidade nos anos 2021, 2022 e 2023 - Município IBGE 290570 (Camaçari)	2021: 5,52% 2022: 3,16% 2023: 2,14%
Painel de Notificações de COVID-19	**Resultado**
QC 01: Óbitos que ocorreram em função do COVID-19	Total: 565
QC 02: Total de pessoas se recuperaram plenamente do COVID-19	Total: 26216
QC 03: Total de pessoas vacinadas pegaram COVID-19 (confirmados em laboratório)	Total: 410
Painel de Saúde Materno-Infantil (SIM - SINASC)	**Resultado**
QC 01: Número de Mulheres que vieram a óbito em Idade Fértil (faixa etária de 10 a 49 anos)	1418
QC 02: Taxa de Mortalidade Infantil por 1000 nascidos vivos - Município IBGE 290570 (Camaçari)	2021: 0,81% 2022: 0,77% 2023: 0,30%
Painel de Notificações x SIM x SINASC	**Resultado**
QC 01: Óbitos com registro no SIM	Total: 439
QC 02: Mortalidade por COVID-19 com registro no SIM	Total: 389
QC 03: Mortalidade por COVID-19 com registro no SIM por ano de nascimento	2020: 158 2021: 277 2022: 9

7. Conclusão e Trabalhos Futuros

O objetivo principal deste trabalho foi realizar a integração semântica entre as bases heterogêneas SIM, SINASC e e-SUS Notifica, utilizando grafos de conhecimento semântico, com o propósito de apoiar gestores do SUS no acesso a informações integradas sobre nascimentos, óbitos e notificações de COVID-19.

Esse objetivo foi alcançado por meio do desenvolvimento da ontologia Ontovid II, construída a partir de uma abordagem híbrida de integração. A solução permitiu responder a questões de competência formuladas pela Secretaria Municipal de Saúde de Camaçari-BA, com resultados validados pela equipe técnica, demonstrando seu potencial de aplicação prática em ambiente produtivo. Assim, pode-se afirmar que a questão de pesquisa que orientou este estudo — "Como integrar semanticamente bases heterogêneas de saúde pública (SIM, SINASC e e-SUS Notifica), de modo a disponibilizar informações integradas que apoiem gestores do SUS na tomada de decisão?" — foi respondida com a construção e validação da ontologia Ontovid II.

Apesar dos resultados alcançados, este estudo apresenta algumas limitações. A solução foi validada apenas com dados do município de Camaçari-BA, o que restringe sua generalização imediata para outros contextos. Além disso, não foi desenvolvido um protótipo de interface para uso direto pelos gestores, o que limita a aplicabilidade imediata em ambiente produtivo. Por outro lado, os resultados evidenciam relevantes implicações práticas: a utilização de grafos de conhecimento semântico mostrou-se eficaz para integrar dados heterogêneos de saúde, reduzindo a complexidade técnica do acesso às informações e oferecendo suporte direto para gestores municipais no acompanhamento de indicadores de mortalidade, natalidade e notificações de COVID-19.

Como trabalhos futuros, pretende-se: i) incorporar outras bases de dados de saúde, como a da Campanha Nacional de Vacinação contra a COVID-19; ii) realizar o alinhamento das ontologias desenvolvidas com ontologias de alto nível, como DOLCE ou BFO; iii) desenvolver um protótipo funcional para apresentar os resultados em ambiente produtivo e; iv) implementar métricas quantitativas, avaliação de desempenho e testes de escalabilidade para que se possa avaliar a robustez da integração sob uso em larga escala. Além disso, pretende-se disponibilizar as ontologias, arquivos de mapeamento e consultas SPARQL em repositório aberto, acompanhados de conjuntos de dados sintéticos anonimizados, a fim de favorecer a reprodutibilidade e a transparência metodológica. Também se planeja expandir os experimentos para outros municípios, explorando a generalização da abordagem em cenários com diferentes perfis e padrões de qualidade de dados.

Agradecimentos

Os autores gostariam de agradecer o suporte financeiro recebido da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB), por meio do projeto "UAware: Uma Plataforma de Comunicação de Crise Ubíqua e Sensível ao Contexto de apoio ao Gerenciamento de Emergências em Cidades e Áreas Industriais"(Termo de Outorga TIC 0008/2015).

Agradecemos também à Prefeitura Municipal de Camaçari, em especial à Secretaria de Saúde e à Diretoria de Planejamento, pelo indispensável acesso aos dados, pelo compartilhamento de conhecimento técnico do domínio e pelo constante apoio durante a realização desta pesquisa.

Declaração sobre o Uso de Inteligência Artificial Generativa

Durante a preparação deste trabalho, os autores utilizaram o ChatGPT (GPT-4, OpenAI) e o GitHub Copilot para apoio no aprimoramento textual, verificação gramatical e ortográfica, bem como para sugestões de código. Após a utilização dessas ferramentas, o conteúdo foi integralmente revisado e editado pelos autores, que assumem total responsabilidade pelo conteúdo da publicação.

Referências

N. Guarino, D. Oberle, S. Staab, What is an ontology?, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 1–17. URL: https://doi.org/10.1007/978-3-540-92673-3_0. doi:10.1007/978-3-540-92673-3_0.

P. Grenon, B. Smith, Foundations of an ontology of philosophy, Synthese 182 (2011) 185–204. doi:10.1007/s11229-009-9658-x.

L. D. de Cerqueira, Uma Abordagem Baseada em Ontologias para Integração Semântica de Sistemas na Camada de Processos, Master's thesis, Universidade Federal do Espírito Santo, 2016.

N. Gomes, S. Franco, L. Salvador, Ontovid - uma abordagem para construção de grafos de conhecimento semântico com enfoque em notificações e Óbitos relacionados ao novo coronavírus (covid-19), 2022, pp. 425–436. doi:10.5753/sbcas.2022.222723.

F. Ekaputra, M. Sabou, E. Serral, E. Kiesling, S. Biffl, Ontology-based data integration in multi-disciplinary engineering environments: A review, Open Journal of Information Systems (OJIS) 4 (2017) 1–26.

S. Flesca, S. Greco, E. Masciari, D. Saccà, A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, volume 31, 2018. doi:10.1007/978-3-319-61893-7.

A. Y. H. Natalya F. Noy, AnHai Doan, Semantic integration (editorials), ai magazine, 2005, p. 26(1). doi:https://doi.org/10.1609/aimag.v26i1.1794.

R. Calhau, R. Falbo, An ontology-based approach for semantic integration, 2010, pp. 111 – 120. doi:10.1109/EDOC.2010.32.

S. P. Gardner, Ontologies and semantic data integration, Drug Discovery Today 10 (2005) 1001–1007. URL: https://www.sciencedirect.com/science/article/pii/S135964460503504X. doi:https://doi.org/10.1016/S1359-6446(05)03504-X.

T. Bagosi, D. Calvanese, J. Hardi, S. Komla-Ebri, D. Lanti, M. Rezk, M. Rodríguez-Muro, M. Slusnys, G. Xiao, The ontop framework for ontology based data access, in: D. Zhao, J. Du, H. Wang, P. Wang, D. Ji, J. Z. Pan (Eds.), The Semantic Web and Web Science, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 67–77.

C. M. Keet, An introduction to ontology engineering, volume v1.5, 2020, pp. 165–175.

D. Calvanese, M. Giese, D. Hovland, M. Rezk, Ontology-based integration of cross-linked datasets, in: M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d'Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, S. Staab (Eds.), The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 199–216.

H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Ubner, Ontology-based integration of information - a survey of existing approaches, Proceedings of the IJCAI'01 Workshop on Ontologies and Information Sharing, Seattle, Washington, USA, Aug 4-5 (2002).

A. Soylu, O. Corcho, E. Simperl, D. Roman, F. Martínez, C. Taggart, I. Makgill, B. Elvesaeter, B. Symonds, H. Mcnally, G. Konstantinidis, Y. Zhao, T. Lech, Towards integrating public procurement data into a semantic knowledge graph, 2018.

D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, R. Rosati, Ontology-Based Data Access and Integration, Springer International Publishing, 2018, p. 2590–2596.

A. B. B. Junior, Um framework baseado em conhecimento de senso comum para sistemas de perguntas e respostas sobre grafo de conhecimento, 2021. URL: https://repositorio.ufrn.br/handle/123456789/45676.

M. Cheatham, C. Pesquita, Semantic data integration. in handbook of big data technologies, Springer International Publishing (2017) 263–305. doi:https://doi.org/10.1007/978-3-319-49340-4.

P. Auceli, R. Berardi, N. Kozievitch, Integração semântica entre dados dos domínios da educação e segurança: um caso em curitiba, in: Anais da XV Escola Regional de Banco de Dados, SBC, Porto Alegre, RS, Brasil, 2019, pp. 141–150. URL: https://sol.sbc.org.br/index.php/erbd/article/view/8487. doi:10.5753/erbd.2019.8487.

R. B. de Oliveira, Utilização de Ontologias Para Busca em Base de Dados de Acórdãos do STF, Master's thesis, Universidade de São Paulo, São Paulo, 2017.

A. Hajmoosaei, S. Abdul-Kareem, An ontology-based approach for resolving semantic schema conflicts in the extraction and integration of query-based information from heterogeneous web data sources, in: Proceedings of the Third Australasian Workshop on Advances in Ontologies - Volume 85, AOW '07, Australian Computer Society, Inc., AUS, 2007, p. 35–43.

L. Zhao, R. Ichise, Ontology integration for linked data, Journal on Data Semantics 3 (2014) 237–254. URL: https://doi.org/10.1007/s13740-014-0041-9. doi:10.1007/s13740-014-0041-9.

G. Xiao, D. Hovland, D. Bilidas, M. Rezk, M. Giese, D. Calvanese, Efficient ontology-based data integration with canonical iris, in: A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, M. Alam (Eds.), The Semantic Web, Springer International Publishing, Cham, 2018, pp. 697–713.

T. Vidal, C. Viktor, S. Avila, R. Mariano, T. Calixto, P. Ivo, J. Filho, A. Brayner, M. Vidal, Uso das tecnologias da web semântica na construção de grafos de conhecimento semântico baseado no enfoque híbrido, 2021.

D. L. N. C. Pereira, Integração semântica das bases de dados do Sistema Único de Saúde: um estudo de caso com o Município de São Paulo, Master's thesis, Universidade de São Paulo, São Paulo, 2019.

M. M. Lima da Cruz, C. Viktor, V. Vidal, N. Arruda Junior, Semanticsus: Um portal semântico baseado em ontologias e dados interligados para acesso, integração e visualização de dados do sus, 2019, pp. 13–18. doi:10.5753/sbcas.2019.6277.

L. Maddalena, F. Baião, Ontocovid: Aplicando sabio para a modelagem conceitual bem fundamentada no domínio da covid-19. (ontocovid: Applying sabio to conceptual modeling well grounded in the covid-19 domain), in: F. Farinelli, R. C. G. Berardi, J. L. Carbonera, D. Schmidt (Eds.), Proceedings of the XIV Seminar on Ontology Research in Brazil (ONTOBRAS 2021) and V Doctoral and Masters Consortium on Ontologies (WTDO 2021), Online, Brazil, November 16-19, 2021, volume 3050 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 259–266. URL: http://ceur-ws.org/Vol-3050/Short5.pdf.

F. L. G. D. Fernantes, Integração de Dados Baseada em Ontologias e Raciocínio Automático: Estudo de Caso com Dados Públicos de Saúde, 2012. URL: https://repositorio.ufpe.br/handle/123456789/10973.

ATHENA: A FAIR approach to publish and evaluate cybersecurity datasets*

Thaisa da S. Hernandez[1,4,*,†], Caroline Duarte Gandolfi[1,†], Pedro Henrique Bulcão[1,†], Luiz Bonino da Silva Santos[2,†], Anderson F. P. dos Santos[1,3,†] and Maria Cláudia Reis Cavalcanti[1,†]

[1] *Instituto Militar de Engenharia, Praça Gen. Tibúrcio 80, Urca, Rio de Janeiro, RJ, 22290-270*

[2] *University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*

[3] *Venturus Centro de Inovação Tecnológica, Av. G. V. di Napoli 1185, Bosque das Palmeiras, Campinas, SP, 13086-530*

[4] *Diretoria de Comunicações e Tecnologia da Informação da Marinha, Rua 1º de Março, 118, Centro, Rio de Janeiro, RJ, 13086-530*

Abstract

The massive increase in the attack surface caused by an exponential volume of data has highlighted the importance of continuous research in the field of cybersecurity, which in turn has become increasingly data-driven. The availability and quality of cybersecurity datasets are therefore fundamental for the reliability of predictions and the implications in innovation in this domain. However, there are numerous challenges regarding the availability of good-quality cybersecurity datasets. This work addresses these challenges by proposing an approach to publish cybersecurity dataset metadata and to assess the quality of these datasets, considering their specific properties. The differential of our approach is the integration of the FAIR (Findable, Accessible, Interoperable, Reusable) principles into the evaluation process. This approach was implemented as a composite of software modules. First, a FAIR Data Point repository was instantiated to publish metadata about cybersecurity datasets. Secondly, the Athena Evaluator module was implemented to analyze the metadata published in the repository based on a set of specific quality metrics and on metrics aligned with the FAIR principles. Additionally, to support the creation and management of different metadata schemas for the various types of cybersecurity datasets, we have also developed an easy-to-use form design tool, named FAIR Data Point metadAta Schema ediTor (FAST), that provides agility and flexibility to the metadata repository platform. Last but not least, we created a metadata schema for network traffic datasets based on the lightweight Athena-o ontology, which provides a semantic basis for describing the properties of these datasets.

Keywords

FAIR principles, dataset evaluation, metadata, information security

1. Introduction

The ever-increasing number of digital threats requires a continuous advance in cybersecurity research and practices, which are becoming increasingly data-driven [1]. In this scenario, cybersecurity datasets play a key role, serving as the basis for training machine learning models, validating intrusion detection systems, analyzing malware, and investigating new vulnerabilities [2]. The availability and quality of these datasets are therefore fundamental to the reliability of predictions [3] and to driving innovation in the field of cybersecurity.

However, there are still a few quality cybersecurity datasets available to be reused [4]. The main concerns about sharing cybersecurity data are the challenges of preserving privacy and standardizing the data publication format [4]. The relative scarcity of cybersecurity datasets is compounded by the

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

[†] These authors contributed equally.

✉ thaisa.silva@ime.eb.br (T. d. S. Hernandez); caroline.gandolfi29@ime.eb.br (C. D. Gandolfi); pedrohgbulcao@ime.eb.br (P. H. Bulcão); l.o.boninodasilvasantos@utwente.nl (L. B. d. S. Santos); anderson@ime.eb.br (A. F. P. d. Santos); yoko@ime.eb.br (M. C. R. Cavalcanti)

0009-0000-8310-796X (T. d. S. Hernandez); 0009-0006-1173-8600 (C. D. Gandolfi); 0009-0008-6666-0104 (P. H. Bulcão); 0000-0002-1164-1351 (L. B. d. S. Santos); 0000-0002-6754-4809 (A. F. P. d. Santos); 0000-0003-4965-9941 (M. C. R. Cavalcanti)

lack of a central registry and inconsistent provenance information. In addition, most cybersecurity datasets are outdated, and much of the information related to attack data is redundant [5]. With regard to the quality of a cybersecurity dataset, there are clear challenges in obtaining, maintaining, and publishing it. Besides, there is a shortage of consistent metrics, and researchers limit themselves to evaluating quality based on the reputation of the authors [5].

These challenges result in a central problem: the lack of a formal procedure to publish metadata and evaluate the quality and reliability of cybersecurity datasets. This work directly addresses this problem by proposing an approach to publish dataset metadata and evaluate the quality of these datasets, considering their specific properties. A differential of our approach is the integration of the FAIR principles [6] into the evaluation process. They provide guidelines for the publication of digital resources such as datasets, in a way that makes them Findable, Accessible, Interoperable, and Reusable [6]. By incorporating the FAIR principles, we not only aim to measure the technical quality of the cybersecurity datasets but also to promote better data management and reuse practices. As a secondary objective, we aim to contribute to increasing the availability of and trust in high-quality cybersecurity datasets.

To achieve these goals, a metadata repository has been implemented to support flexible schemas, adapted to the specific properties of the various types of cybersecurity datasets. Quality evaluation is carried out by the Athena Evaluator software, which analyzes the metadata published in the repository based on a set of specific quality metrics and also metrics aligned with the FAIR principles. To support the creation and management of these metadata schemas, we have also developed a lightweight ontology, which provides a semantic basis for describing the properties of cybersecurity datasets.

This article is organized as follows: Section 2 presents related work. Section 3 presents the Athena approach. Section 4 presents the implementation of this approach in the context of network traffic datasets. In Section 5, we present a case study on the evaluation of the CIC-DDoS2019 dataset. Finally, in Section 6, we conclude the paper and discuss the next steps in our research.

2. Related Works

The related work was organized to cover research that evaluates the quality of cybersecurity datasets and research that focuses on FAIR data management. Data quality assessment is a well-established field of research in several areas [7], but its specific application to cybersecurity datasets presents unique challenges related to the dynamic, heterogeneous, and sensitive nature of these data [8]. Gharib et al. [9] conducted a study of existing cybersecurity datasets between 1998 and 2016, and presented an evaluation framework for cybersecurity datasets with eleven proposed criteria: complete network configuration, complete traffic, labeled dataset, complete interaction, complete capture, available protocols, attack diversity, anonymity, heterogeneity, feature set, and metadata. These eleven criteria are evaluated according to a weight that can be defined on the basis of the organization's request or the type of Intrusion Detection System (IDS) selected for the test. In Sharafaldin et al. [10], a specific cybersecurity dataset was developed, and the quality of this dataset was compared to other synthetically generated datasets. This comparison was based on the eleven criteria proposed by Gharib et al. [9] in his framework. However, although the evaluation structure proposed by Gharib et al. is quite complete, containing a range of quality criteria and a quantitative approach to evaluating these criteria, there is a gap related to checking the timeliness of the dataset. In Ring et al.[11], a survey focused on cybersecurity datasets was carried out, where a collection of fifteen properties was established as a basis for identifying and comparing these datasets. These properties cover a range of criteria and are grouped into five categories: general information, nature of the data, volume of data, recording environment, and evaluation, but do not create a scoring structure to evaluate these criteria. Furthermore, despite agreeing with the FAIR Principles, this work does not go into these principles.

Regarding related work on FAIR data management in the field of cybersecurity datasets, Raza et al.[12] uses the FAIR Principles as a framework for data management and evaluation, reinforcing the importance of making data Findable, Accessible, Interoperable, and Reusable. The article proposes a

Table 1
Related works

	[9]	[10]	[12]	[11]	[13]	[14]	[4]	**This work**
Cybersecurity Dataset quality	x	x		x	x			x
FAIR Principles adequacy			x	x		x	x	x
Customizable cybersecurity metadata						x		x
Lightweight ontology							x	x

methodology for developing and evaluating fair-compliant datasets, although it is in a different domain of cybersecurity, focused on Large Language Models (LLMs). Silva et al. [4] proposed an approach to support cybersecurity dataset publishing for machine learning tasks following FAIR principles and involving, among others, anonymization and preprocessing of data. This approach addresses the limited availability of cybersecurity datasets, providing an environment to facilitate and motivate the creation of these datasets for publication. However, the emphasis of the approach is on generating higher-quality data in line with the FAIR principles, rather than covering a process of evaluating the dataset prior to its publication. The research carried out by Göbel et al. [13] has a focus on the creation and optimization of datasets in the context of cybersecurity, with an emphasis on digital forensics. It addresses the challenges and best practices for creating high-quality datasets, although it does not explicitly address the fairness of datasets. Mombelli et al. [14] addresses the application of the FAIR Principles and metadata quality in the field of digital forensics. The paper evaluates metadata completeness and compliance with the FAIR Principles in 212 datasets from NIST's Computer Forensic Reference Dataset Portal (CFReDS). The results indicate deficiencies in metadata quality and the need for better data management standards. Providing important insights into the ongoing need to improve metadata management in cybersecurity datasets.

Unlike the aforementioned works, this paper combines data management and quality assessment in the field of cybersecurity. The Athena approach was based on three fundamental pillars: a customizable FAIR Data Point repository, a lightweight support ontology called Athena-o, and the Athena Evaluator software for evaluating specific cybersecurity metrics and FAIR metrics. By incorporating FAIR principles as a new dimension of quality assessment, we aim not only to measure the technical quality of cybersecurity datasets but also to promote best practices in data management and reuse. Table 1 summarizes the differential of the Athena approach.

3. Athena Approach

The Athena approach aims to evaluate the quality of cybersecurity datasets and to help promote better data management and sharing practices. To this end, we integrate the analysis of the intrinsic properties of cybersecurity datasets with the evaluation of their compliance with the FAIR principles. Our approach is extensible, capable of adapting to the diversity of datasets in the cybersecurity domain, such as network traffic and malware datasets. The Athena approach is based on three fundamental pillars: a customized metadata repository, a lightweight support ontology called Athena-o and the Athena Evaluator software. Figure 1 gives an overview of the Athena approach, and each stage is detailed below.

Quality evaluation begins with the publication of the dataset through a set of descriptive metadata, stored in our customized repository. This repository has been implemented to be flexible, allowing the definition of specific metadata schemes for different types of cybersecurity datasets. The central idea is that quality is not an absolute concept, "quality data must be intrinsically good, contextually appropriate for the task and clearly represented to the data consumer" [15]. In the step **Select a Cybersecurity Dataset Type**, the person responsible for publishing the cybersecurity dataset metadata, in this approach called the Publisher, selects the most appropriate metadata schema to describe their type of cybersecurity dataset.

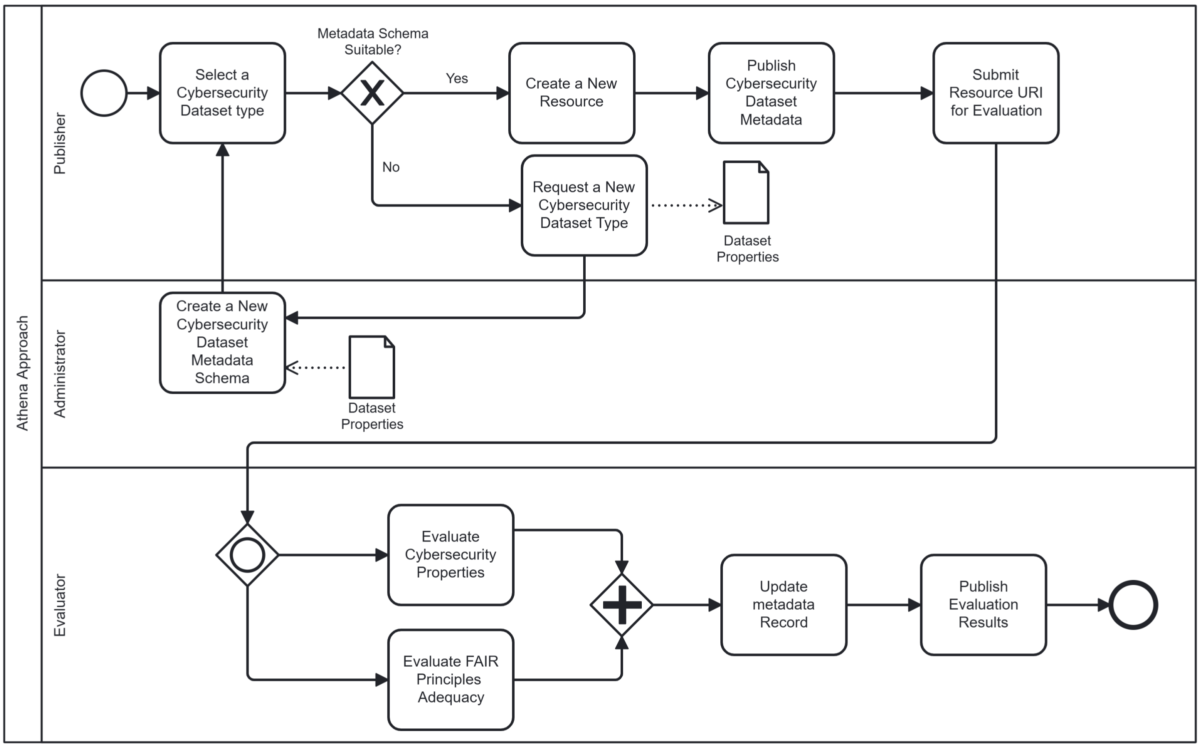

Athena Approach
Publisher
Administrator
Evaluator
Select a Cybersecurity Dataset type
Metadata Schema Suitable?
Yes
No
Create a New Resource
Publish Cybersecurity Dataset Metadata
Submit Resource URI for Evaluation
Request a New Cybersecurity Dataset Type
Dataset Properties
Create a New Cybersecurity Dataset Metadata Schema
Dataset Properties
Evaluate Cybersecurity Properties
Evaluate FAIR Principles Adequacy
Update metadata Record
Publish Evaluation Results

Figure 1: Athena approach - publication and evaluation process

The Administrator plays the role of the person responsible for creating metadata schemas. If there is no suitable metadata schema, the step **Request a New Metadata Schema** can be triggered and the Administrator can **Create a New Cybersecurity Dataset Metadata Schema** informing the necessary properties. The definition of these schemas is supported by our lightweight ontology, which provides semantic relationships to describe the properties consistently. Moreover, this task should provide a form editor facility, which allows users to create these schemas, facilitating the extensibility of the approach. Details on the implementation of the metadata repository, the lightweight ontology, and the form editor will be presented in section 4.

In our approach, a dataset is registered in the repository as a resource. So, in the next step **Create a new Resource**, a new digital resource is created containing metadata records from a specific dataset according to the selected metadata schema. Once the resource has been created, in the step **Publish Cybersecurity Dataset Metadata** the Cybersecurity Dataset metadata is recorded and made available for evaluation.

When the Publisher **Submit Resource URI for Evaluation**, the Evaluator software starts interacting with the metadata repository with the URI of a resource provided. At this stage, the dataset will undergo two types of evaluation. In the step **Evaluate Cybersecurity Properties**, the dataset will be evaluated based on a set of metrics defined on the basis of specific cybersecurity properties. These properties, from a general perspective, can cover aspects such as data timeliness and relevance. The selection and weighting of the metrics can be adjusted depending on the type of dataset, giving flexibility to the process. In addition, in the stage **Evaluate FAIR principles adequacy** the dataset is subjected to maturity tests using FAIR Metrics[1] to evaluate its level of compliance with the FAIR principles. The results of the evaluations are published together with the other metadata records.

[1] https://github.com/FAIRMetrics/Metrics/

4. Implementation

To implement the Athena approach, a FAIR Data Point repository[2] was customized to support a specific metadata schema, adapted to the specific properties of the various types of cybersecurity datasets. The FAIR Data Point follows the Data Catalog Vocabulary (DCAT)[3], and one of its main differential characteristics is its flexibility, i.e., it may be customized to describe different types of digital objects, which are defined as sub-classes of DCAT Resource. This work takes advantage of this feature, using the inheritance of DCAT's general properties and focusing only on specific features of cybersecurity datasets.

Although the Athena approach aims to cover a variety of cybersecurity datasets, its initial implementation and validation focused on network traffic datasets. In the context of cybersecurity, network traffic datasets have specific characteristics that need to be considered, such as the year in which the traffic was generated, which is different from the year in which the dataset itself was published; the incidence of malicious traffic and the corresponding types of attack; whether the traffic was labeled or not; and the type of network on which the traffic was generated. Ring et al. [11] summarized this set of properties into five categories: general information (year of traffic creation, public availability, normal traffic, attack traffic), nature of the data (metadata, format, anonymity), data volume (count and duration), recording environment (traffic type, network type, complete network) and evaluation (predefined, balanced, labeled divisions).

In this section, we first describe the lightweight ontology named Athena-o (subsection 4.1), which reused concepts of existing ontologies, conforming to the Interoperability principle. From this ontology, we derived the Athena metadata schema (subsection 4.2), which included the properties already mentioned by Ring et al. [11], extending the DCAT schema. Finally, the Athena Evaluator (subsection 4.3) was implemented using the FAIR metrics API[1], which already implements metrics to evaluate datasets concerning the FAIR principles. However, we implemented new specific metrics to evaluate the network traffic datasets, based on the specific properties defined in the Athena metadata schema.

4.1. Athena-o

The Athena-o lightweight ontology, shown in Figure 2, was developed to provide a semantic basis and interoperability in the selected properties that describe cybersecurity datasets. This ontology defines new concepts, as well as reuses existing classes from well-known ontologies and vocabularies, such as Dublin Core (DC)[4], TOUCAN Ontology (ToCo)[5], Unified Cyber Ontology (UCO)[6] and National Institute of Standards and Technology (NIST) glossary[7].

By extending the **DCAT Dataset** concept, Athena-o reuses the already well-established properties relating to dataset metadata such as *dcterms:format* and *dcat:byteSize*. In this article, we focus on the specific features of cybersecurity datasets. Athena-o introduces the **Cybersecurity Dataset** concept (*at:CybersecurityDataset*), which specializes the **DCAT dataset** concept (*dcat:Dataset*), of which, in turn, the **Network Traffic Dataset** (*at:NetworkTrafficDataset*) concept is specialized. Furthermore, specific properties have been defined for the **Network Traffic Dataset** concept conforming to the properties defined by Ring et al. [11], such as the year of traffic creation (*time:yearOfTrafficCreation*), and the kind of traffic (*at:kindOfTraffic*), whose values may be *real*, *emulated*, or *synthetic*. The **Traffic** concept inherits from the **UCO Network Flow** concept (*uco:NetworkFlow*), which can be specialized into two concepts: **Normal network traffic** (*at:NormalTraffic*) and **Anomalous network traffic** (*at:AttackTraffic*). The **Attack Traffic** concept is connected to the **Attack Type** concept (*nist:attack*), reused from the NIST Glossary, through the property *at:isClassifiedBy*. This concept has two properties that represent the attackers' IP (*at:AttackerIP*) and the victims' IP (*at:VictimIP*).

[2] https://app.fairdatapoint.org/

[3] https://www.w3.org/TR/vocab-dcat-3/

[4] https://www.dublincore.org/

[5] https://github.com/QianruZhou333/toco_ontology/

[6] https://unifiedcyberontology.org/

[7] https://csrc.nist.gov/glossary/

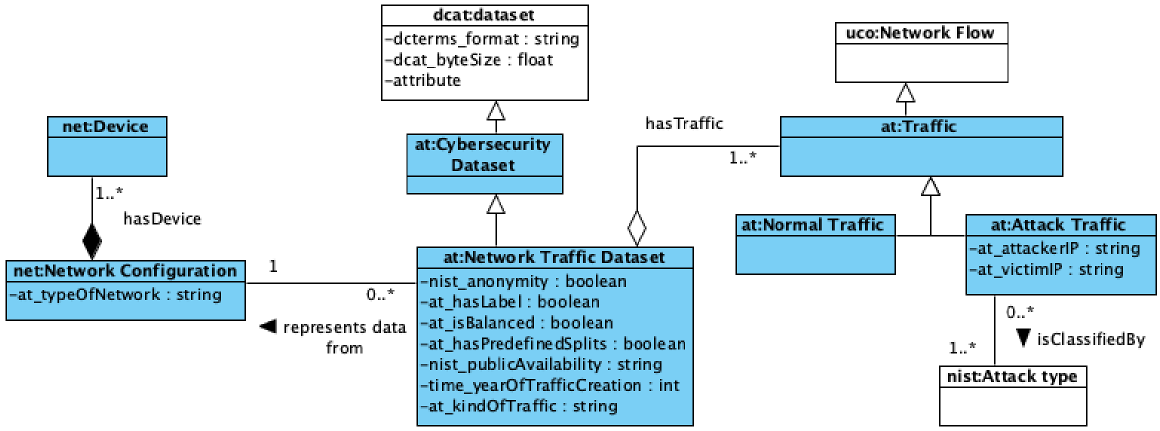


Figure 2: Athena-o: a lightweight ontology for the Athena approach

According to Ring et al., a dataset description must include the network configuration through which the traffic flowed. Thus, Athena-o reuses the **Physical infrastructure** (*net:PhysicalInfrastructure*) and **Device** (*net:Device*) concepts from the Toco Ontology. The former includes a property that represents the type of network from which the data in a dataset was collected (*at:typeOfNetwork*), and the latter represents the devices that are part of a network infrastructure. In addition, *nist:hasPublicAvailability* and *nist: Anonymity* properties represent the availability and anonymization of a dataset, respectively. Finally, *at:hasPredefinedSplits*, *at:hasLabel*, and *at:isBalanced* properties represent metadata that are useful for performing effective machine learning tasks. These properties provide respectively information if a dataset includes predefined subsets for training and evaluation, if datasets are labeled or not and if datasets are balanced with respect to their class labels.

The applicability of the approach to other types of cybersecurity datasets (e.g., malware) is possible through the extension of the Athena-o ontology, in which case a new class would be added as a subclass of Cybersecurity Datasets. This extension of the ontology and the subsequent creation of a metadata schema are performed by the Administrator, based on the new properties submitted by the publisher, as described in Section 3.

4.2. Athena Metadata Schema

The Athena metadata schema is expressed in RDF (Resource Description Framework) using the Shapes Constraint Language (SHACL) [16] and the Data Shapes Vocabulary (DASH)[8]. The former is rich in establishing constraints for validating the schema instantiations, while the latter is an extension of SHACL with new constraints and target types, and also includes components to fix constraint violations. Moreover, SHACL includes constructs such as *sh:order* and *sh:group* that can aid in the construction of form layouts, and DASH also includes constructs that are particularly useful for form configuration, such as *dash:TextFieldEditor*.

Athena-o guided the creation of the corresponding metadata schema, but some simplifications were made. The choice for the enrichment of the schema with SHACL and DASH languages was required by the FAIR Data Point implementation. Listing 1 shows a fragment of the Athena metadata schema created for describing the Network Traffic datasets. Note that it begins with the declaration of the *DatasetShape* element that has the *dcat:Dataset* as its target class. For simplification reasons, besides the dataset element, all the other elements of Athena-o were mapped into properties associated to the dataset element. The DASH constructs inform the FAIR Data Point user interface elements, so it can organize and configure the properties in a form for capturing metadata values.

For example, the *publicAvailability* attribute is defined as a drop down menu (*Select field*), which guides the user in choosing one of the pre-defined options. According to Athena-o, a Network Traffic

[8] https://datashapes.org/dash/

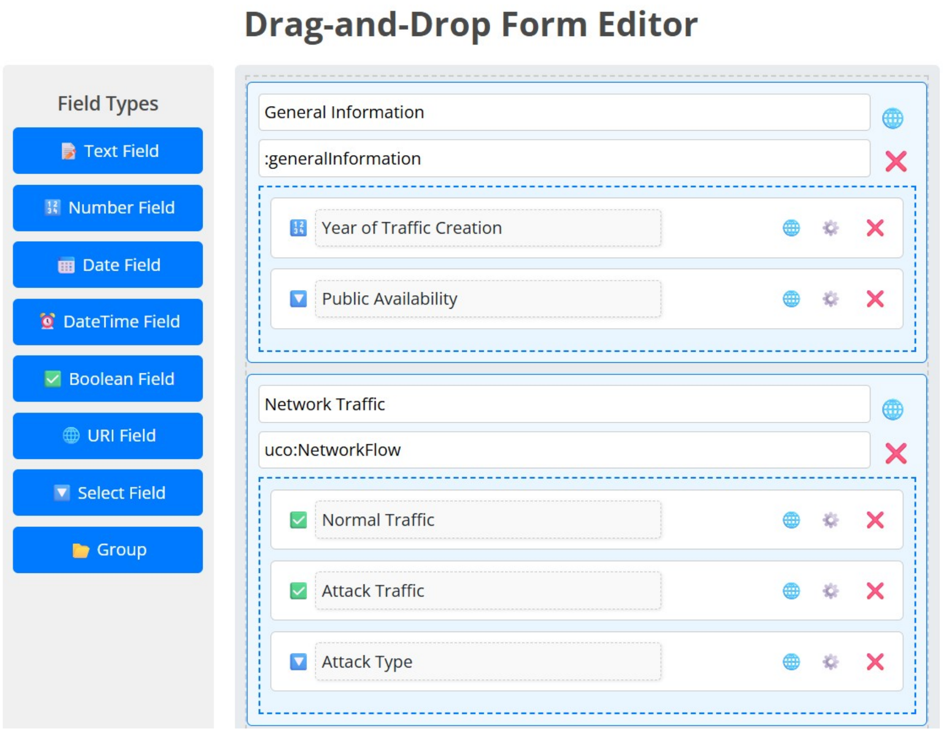
Dataset contains Traffic that may be Normal or Attack Traffic. Note that, to describe the dataset metadata, there is no need to represent its content, but it is important to indicate if it includes Normal or Attack traffic. Thus, *Attack/Normal traffic* properties are defined as boolean datatypes. Similarly, the *Attack type* is also defined as a property associated directly with the dataset, indicating what types of attack it includes.

Listing 1: Metadata Schema Fragment for Network Traffic Datasets

```
:DatasetShape a sh:NodeShape ;
    sh:targetClass dcat:Dataset ;
    sh:property [
        sh:path time:year ;
        sh:name "Year of Traffic Creation" ;
        sh:datatype xsd:integer ;
        dash:editor dash:TextFieldEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group :generalInformation ;
        sh:order 0 ;
    ] ;
    sh:property [
        sh:path at:publicAvailability ;
        sh:name "Public Availability" ;
        sh:datatype xsd:string ;
        sh:in ( "No" "On request (o.r.)" "Yes" ) ;
        dash:editor dash:EnumSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group :generalInformation ;
        sh:order 1 ;
    ] ;
    sh:property [
        sh:path at:NormalTraffic ;
        sh:name "Normal Traffic" ;
        sh:datatype xsd:boolean ;
        dash:editor dash:BooleanSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group uco:NetworkFlow ;
        sh:order 2 ;
    ] ;
    sh:property [
        sh:path at:AttackTraffic ;
        sh:name "Attack Traffic" ;
        sh:datatype xsd:boolean ;
        dash:editor dash:BooleanSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 1 ;
        sh:group uco:NetworkFlow ;
        sh:order 3 ;
    ] ;
    sh:property [
        sh:path nist:attack ;
        sh:name "Attack Type" ;
        sh:datatype xsd:string ;
        dash:editor dash:InstancesSelectEditor ;
        dash:viewer dash:LiteralViewer ;
        sh:minCount 0 ;
        sh:maxCount 100 ;
        sh:group uco:NetworkFlow ;
```

```
            sh:order 4 ;
    ] ;
```

Finally, we highlight that the created schema is easily extended or adapted to other types of cybersecurity datasets. A user-friendly form design tool, named FAIR Data Point metadAta Schema ediTor (FAST), was implemented to automate the schema generation. It allows schema designers to configure a user interface form by dragging and dropping interface components into a canvas in a visual way. Then, the form is automatically transformed into a SHACL/DASH specification of the schema, which in turn is the input to the FAIR Data Point schema configuration. Figure 3 shows an example of the FAST interface tool, where, for example, the *Attack Traffic* property is defined with a Boolean field type. While adding all properties and their respective field types, the SHACL/DASH code can be viewed and edited.



General Information
:generalInformation
Year of Traffic Creation
Public Availability
Network Traffic
uco:NetworkFlow
Normal Traffic
Attack Traffic
Attack Type

Figure 3: Form editor for Network Traffic datasets metadata schema.

4.3. Athena Evaluator

Athena Evaluator is an application developed in Python whose function is to evaluate cybersecurity datasets in two aspects: the first aspect relates to the intrinsic properties of these specific types of datasets, described through a metadata schema, and the second relates to the compliance of these datasets with the FAIR Principles. To do this, Athena Evaluator interacts with the FAIR Data Point API[9] and executes the evaluation metrics based on the published metadata of a given dataset.

In our implementation with focus on Network Traffic datasets, Athena Evaluator applies the quality assessment metrics based on the values of the following metadata: Year of Traffic Creation, Public availability, Normal Traffic, Attack Traffic, Anonymity, Complete Network, Predefined Splits, Balanced, Labeled according to the metadata schema.

Since new attack scenarios emerge every day, the age of a cybersecurity dataset plays a very important role [11]. Older datasets may not fully reflect the risks that exist today, since attacks have new variants launched all the time. To evaluate timeliness, the logic *Fuzzy* [17] is used to define the degree of

[9] https://app.fairdatapoint.org/swagger-ui/index.html/

pertinence of the year of creation of the data set in “Old”, “Medium” and “Recent” categories. For the purpose of this research, we established a specific range of time covers from 1998 to nowadays (2025) because the relevance of the datasets generated in this period with the following intervals: Old [1998 to 2007], Medium [2003 to 2019] and Recent [2016 to 2025]. The pertinence of the year the dataset traffic was created to one of the sets is calculated using a triangular pertinence function [17]. Regarding the anonymization metric, the problems of compromised privacy occur when the payload is not encrypted in a dataset with real traffic. So, most datasets have their payloads removed or anonymized, which decreases the usefulness of the dataset but maintains the privacy of the information [9]. Datasets with synthetic or emulated traffic do not suffer from this issue and can keep this information available. Therefore, the evaluation of this metric is directly related to the type of traffic in the dataset, which can have three values: Real, Emulated, and Synthetic. If a dataset has a real traffic type, it means that the data needs to be anonymized; otherwise, if the traffic type is emulated or synthetic, it makes no sense for the data to be anonymized.

Moreover, the dataset is evaluated based on the presence of the following properties: Public availability, Normal and Attack Traffic, Complete Network, Predefined Splits, Balanced, Labeled according to the metadata schema. Finally, the relevance of a dataset is evaluated by a metric that weights the number of its citations (obtained through its DOI). In this metric, the number of citations is attenuated and correlated with the score assigned to the year in which the traffic was created. This approach helps that a dataset’s historical popularity does not overshadow the usage-based relevance of more recent datasets. Concerning the FAIR principles, Athena Evaluator implements a selected set of FAIR metrics[1]. The score for each sub-principle is the average of the corresponding FAIR metrics. The compliance of the evaluated dataset with each of the principles is as follows:

Findable: Metrics are used to verify the existence of globally unique and persistent identifiers associated with the dataset in order for them to be found and resolved by computers. Globally unique means that the identifier is guaranteed to refer unambiguously to exactly one resource in the world, and persistence refers to the requirement that this globally unique identifier is never reused in another context and continues to identify the same resource, even if that resource no longer exists (F1) [18]. In addition, metrics are used to verify the richness of the metadata description (F2). According to Jacobsen et al. [18], it is hard to generally define the minimally required “richness” of this metadata, except that the more generous it is, both for humans and computers, the more specifically findable it becomes in refined searches. Furthermore, the principles (F3) metadata clearly and explicitly include the identifier of the data it describes, and (F4) metadata are registered or indexed in a searchable resource are also evaluated.

Accessible: One of the main objectives of identifying a digital resource is to simultaneously provide the ability to retrieve the record of that digital resource, in a given format, using a clearly defined mechanism: thus, retrievability is a facet of FAIR accessibility [18]. In this case, a set of metrics is used to check the level of recoverability of the data, including authentication/authorization protocols if necessary (A1.1 and A1.). In addition, the FM-A2 metric is used to verify that metadata is accessible, even when the data are no longer available (A2). It is important that consumers have, at the very least, access to high-quality metadata that describes those resources sufficiently to minimally understand their nature and their provenance, even when the relevant data are not available anymore. There is a continued focus on keeping relevant digital resources available in the future [18].

Interoperability: Achieving a “common understanding” of digital resources through a globally understood “language” for machines is the purpose of principle I1. To evaluate this principle, we used the FAIR Metrics to verify the use of a knowledge representation language, vocabularies and ontologies (I1 and I2). In addition, references to other related resources are included in order to verify that the knowledge representing one resource is linked to that of other resources to create a significantly interconnected network of data and services (I3) [18].

Reusable: Digital resources and their metadata must always, without exception, include a license that describes under what conditions the resource can be used, even if it is “unconditional”. Here, metrics are used to verify the presence of a clear and accessible license (R1.1) and a detailed description of the provenance of the dataset (R1.2).

5. Case Study

For the case study, we selected CIC-DDoS2019[10] because it is widely recognized for intrusion detection research, especially for Distributed Denial of Service (DDoS) attacks, contains a wide variety of DDoS attacks in real time and is used by researchers to find the best characteristics and the best model to detect this type of attack with minimal execution time and cost [19]. For this dataset, we collected the metadata needed to populate our FAIR Data Point repository, using the support of the schema defined for network traffic datasets (Section 4.2). We then submitted the dataset for evaluation by the Athena Evaluator software. Figure 4 shows the metadata collected and published according to the created metadata schema and Figure 5 shows the results of the evaluations carried out by Athena Evaluator.

In Figure 4, we point out that the Network Traffic Datasets metadata schema is informed using the *conformsTo* predicate of the Dublin Core Terms [20]. In addition, metadata from the DCAT Resources and Datasets classes, such as *dcterms:license* and *dcterms:rights* are inherited to compose, together with the Network Traffic Datasets schema, the metadata records of the CIC-DDoS2019 dataset.

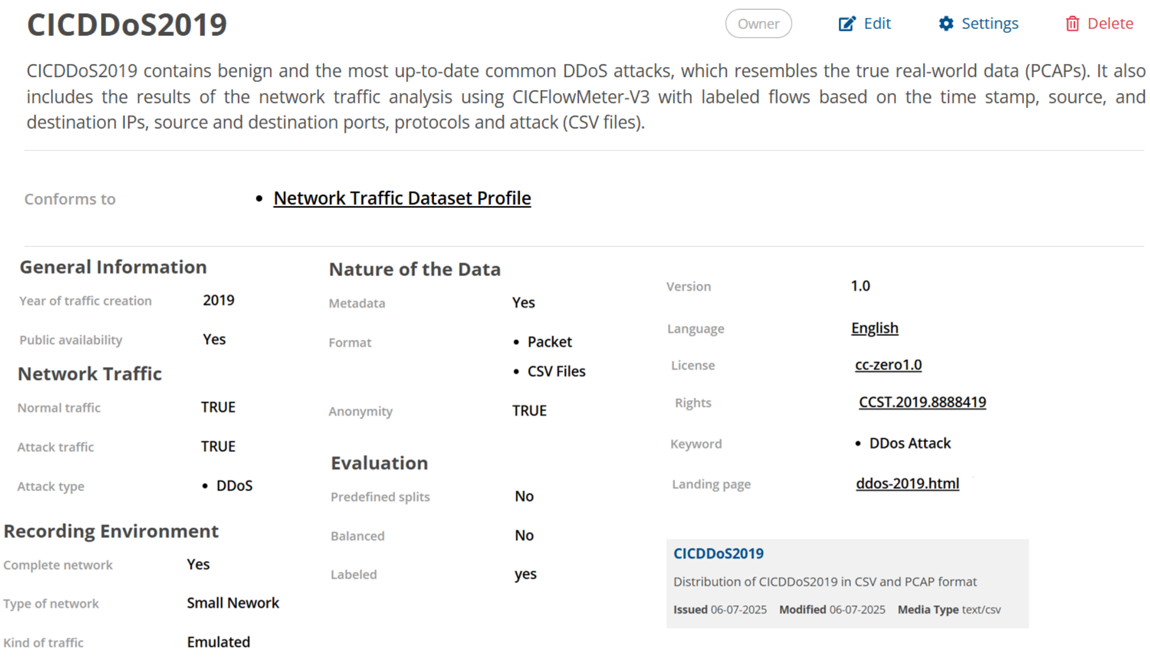
In the first part of the evaluation, metadata for the year of traffic creation, public availability, normal traffic, attack traffic, metadata, anonymity, complete network, predefined splits, labeled and balanced are collected by Athena Evaluator through the FAIR Data Point API[9] and submitted for evaluation according to the specific metrics detailed in Section 4.3. The degree of pertinence of the year of traffic creation of the CIC-DDoS2019 dataset in “Old”, “Medium” and “Recent” categories was calculated using a triangular pertinence function, receiving a higher score for having a higher degree of membership to the "Recent" set. In addition, the dataset is publicly available, contains benign traffic as well as more up-to-date DDoS attacks (DNS, SNMP, NTP, WebDDoS, MSSQL, UDP, LDAP, NetBIOS, SSDP, PortScan, UDP-Lag, and SYN), has a complete network configuration, and makes a good amount of metadata available to the community. Regarding the anonymity metric, since it is a dataset that contains an emulated traffic type, anonymization is not necessary. Furthermore, since it is a labeled dataset, it received the maximum score in this metric. On the other hand, because it is not balanced and does not contain predefined subsets, it did not score in these categories. Finally, its relevance was calculated considering the number of citations and the age score of the dataset.

In the second part of the evaluation, Athena Evaluator assessed the conformity of the CIC-DDoS2019 dataset to the FAIR principles, focusing on its metadata published in the FAIR Data Point repository. The CIC-DDoS2019 dataset performed excellently in the evaluations regarding the principles F1, F2, F3, and F4 due to its rich metadata description and a globally unique and persistent identifier through its DOI. Regarding the Accessibility principle, using HTTP as a communication protocol and publishing its metadata in the FAIR Data Point, which allows for an authentication and authorization procedure when necessary, enabled a good score in these principles (A1.1 and A1.2). Furthermore, the metadata records are available in RDF format, contributing to the principle (I1), and to the use of "vocabularies" such as Dublin Core[4], ToCo[5], UCO[6] and NIST glossary[7] (I2 and I3). For this last test, two metrics were used, in which any Linked Data found was tested for the resolution of a subset of properties (predicates) present and whether these are handled for other Linked Data, failing only the latter. Finally, the dataset was evaluated concerning a clear and accessible data usage license through the *dcterms:license* and *dcterms:rights* (R1.1) and if associated with detailed provenance (R1.2) metadata through, for example, the *dcterms:publisher*, *dcat:contactPoint*, *dcat:landingPage* metadata present in the FAIR Data Point.

The aim is not to score on all the principles, but to encourage the community to provide more accessible, interoperable, and reusable datasets for the advancement of cybersecurity research. By evaluating cybersecurity datasets from this perspective, our study not only contributes to understanding the quality of this specific dataset but also exemplifies the practical application of the FAIR principles for promoting open science and data reuse in a cybersecurity context, encouraging their adoption in future datasets. The Athena Evaluator code is available at Github[11] for evaluation by the community and reproduction of the results presented here.

[10] https://www.unb.ca/cic/datasets/ddos-2019.html/

[11] https://github.com/comp-ime-eb-br/S2C2-IME/tree/main/deliverables/AthenaEvaluator/



CICDDoS2019

CICDDoS2019 contains benign and the most up-to-date common DDoS attacks, which resembles the true real-world data (PCAPs). It also includes the results of the network traffic analysis using CICFlowMeter-V3 with labeled flows based on the time stamp, source, and destination IPs, source and destination ports, protocols and attack (CSV files).

Conforms to
- Network Traffic Dataset Profile

General Information

Year of traffic creation	2019
Public availability	Yes

Network Traffic

Normal traffic	TRUE
Attack traffic	TRUE
Attack type	• DDoS

Recording Environment

Complete network	Yes
Type of network	Small Nework
Kind of traffic	Emulated

Nature of the Data

Metadata	Yes
Format	• Packet • CSV Files
Anonymity	TRUE

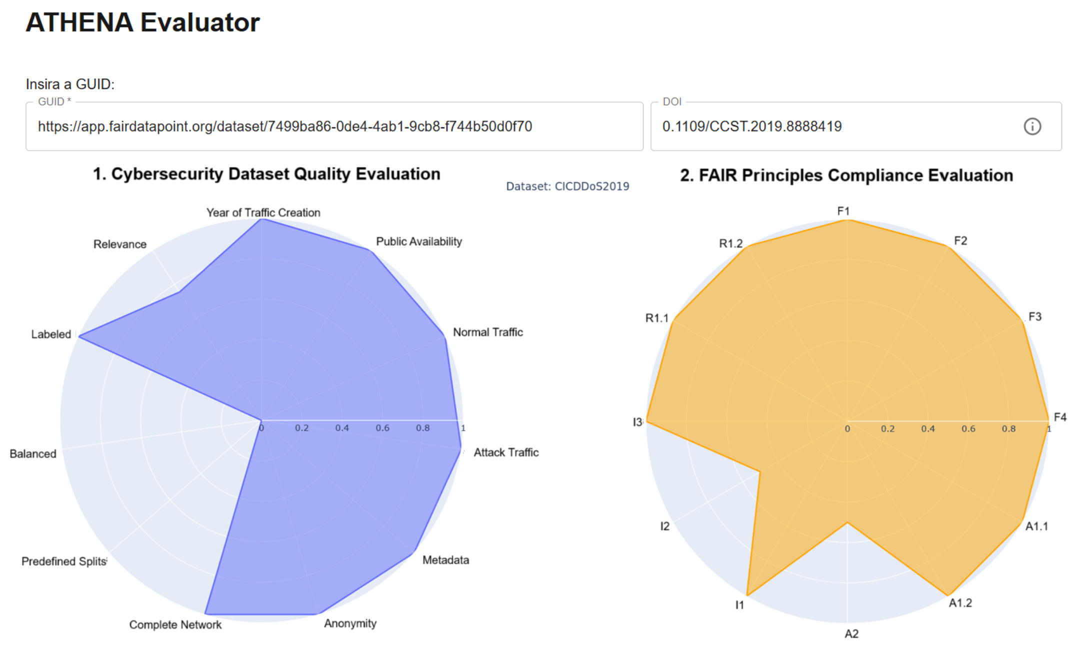
Evaluation

Predefined splits	No
Balanced	No
Labeled	yes

Version	1.0
Language	English
License	cc-zero1.0
Rights	CCST.2019.8888419
Keyword	• DDos Attack
Landing page	ddos-2019.html

CICDDoS2019
Distribution of CICDDoS2019 in CSV and PCAP format
Issued 06-07-2025 **Modified** 06-07-2025 **Media Type** text/csv

Figure 4: FAIR Data Point repository with a specific metadata schema for cybersecurity datasets.



ATHENA Evaluator

Insira a GUID:
GUID *
https://app.fairdatapoint.org/dataset/7499ba86-0de4-4ab1-9cb8-f744b50d0f70
DOI
0.1109/CCST.2019.8888419

1. Cybersecurity Dataset Quality Evaluation

Dataset: CICDDoS2019

2. FAIR Principles Compliance Evaluation

Figure 5: Athena Evaluator with the result of the evaluation of the CICIDS2019 dataset in relation to its specific properties and its compliance with the FAIR principles.

6. Conclusion

This article presented an approach to publish dataset metadata and evaluate the quality of these datasets, considering their specific properties and integration of the FAIR principles into the evaluation process. To this end, the Athena approach was based on three fundamental pillars: a customized FAIR Data Point repository, a lightweight support ontology called Athena-o and the Athena Evaluator software. The FAIR Data Point has been implemented to support flexible metadata schemas, adapted to the specific properties of the various types of cybersecurity datasets. Quality evaluation was carried out by the Athena Evaluator software, which analyzes the metadata published in the repository based on a set of specific quality metrics and also metrics aligned with the FAIR principles. To support the creation and management of these metadata schemas, we have also developed a lightweight ontology, which provides a semantic basis for describing the properties of cybersecurity datasets. We present a case study evaluating the CIC-DDoS2019 dataset, demonstrating the viability of integrating specific properties with FAIR principles, thus structuring a systematic approach with a formal procedure for evaluating cybersecurity datasets. The FAIR Data Point implementation is flexible and can be extended to accommodate new properties and other types of cybersecurity datasets.

By assessing specific properties of cybersecurity datasets as well as potential areas for improvement from a metadata perspective, we provide guidance for researchers involved in creating new datasets. Furthermore, by assessing cybersecurity datasets from this perspective, our study not only contributes to the understanding of the quality of this dataset but also exemplifies the practical application of the FAIR principles to promote open science and data reuse in a cybersecurity context, encouraging their adoption in future datasets. The goal is not to score on all principles, but to encourage the community to provide more findable, accessible, interoperable, reusable, and higher-quality datasets to advance cybersecurity research.

This study focused on the evaluation of dataset quality based on its metadata and the FAIR principles, without delving into the performance of models. For future work, we suggest carrying out a comparative analysis of the impact of the quality characteristics of the evaluated datasets on the performance of different intrusion detection algorithms. In addition, applying this evaluation methodology to other network security datasets could further enrich the understanding of data quality in the area. Finally, we intend to conduct empirical studies that provide further evidence of the applicability of our approach across diverse scenarios and perform a comparative evaluation of other automated frameworks, thereby allowing for a more comprehensive understanding of their performance and the potential advantages of our approach. In this paper, we briefly describe the metrics used to evaluate the datasets. A detailed description will be provided in a future publication.

7. Acknowledgments

This research has been partially supported by CAPES ("Coordenação de Aperfeiçoamento de Pessoal de Nível Superior", in Portuguese). Also, it has been funded by FINEP/DCT/FAPEB (ref.: 2904/20 contract no 01.20.0272.00) under the “System of Systems of Command and Control” project (“Sistema de Sistemas de Comando e Controle”, in Portuguese). Generative AI was used to improve English writing in a few specific points of the text.

8. Declaration on Generative AI

During the preparation of this work, the authors used DeepL for text translation and ChatGPT-5 for citation management. Also, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

[1] M. Zheng, H. Robbins, Z. Chai, P. Thapa, T. Moore, Cybersecurity research datasets: Taxonomy and empirical analysis, in: 11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18), USENIX Association, Baltimore, MD, 2018.

[2] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, Survey of intrusion detection systems: Techniques, datasets and challenges, Cybersecurity 2 (2019) 1–22.

[3] M. Macas, C. Wu, W. Fuertes, A survey on deep learning for cybersecurity: Progress, challenges, and opportunities, Computer Networks 212 (2022) 109032.

[4] M. L. e. Silva, K. de Faria Cordeiro, M. C. Cavalcanti, Sec4ml: An approach to support cybersecurity data publishing for machine learning tasks, in: 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), 2021, pp. 226–235. doi:10.1109/EDOCW52865.2021.00053.

[5] A. Kenyon, L. Deka, D. Elizondo, Are public intrusion datasets fit for purpose? characterising the state of the art in intrusion event datasets, Computers & Security 99 (2020) 102022.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 1–9.

[7] O. Reda, N. C. Benabdellah, A. Zellou, A systematic literature review on data quality assessment, Bulletin of Electrical Engineering and Informatics 12 (2023) 3736–3757.

[8] J. Zhao, M. Shao, H. Wang, X. Yu, B. Li, X. Liu, Cyber threat prediction using dynamic heterogeneous graph learning, Knowledge-Based Systems 240 (2022) 108086.

[9] A. Gharib, I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, An evaluation framework for intrusion detection dataset, in: 2016 International Conference on Information Science and Security (ICISS), IEEE, 2016, pp. 1–6.

[10] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, et al., Toward generating a new intrusion detection dataset and intrusion traffic characterization., ICISSp 1 (2018) 108–116.

[11] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, A. Hotho, A survey of network-based intrusion detection data sets, Computers & Security 86 (2019) 147–167.

[12] S. Raza, S. Ghuge, C. Ding, E. Dolatabadi, D. Pandya, Fair enough: Develop and assess a fair-compliant dataset for large language model training?, Data Intelligence 6 (2024) 559–585. doi:10.1162/dint_a_00255.

[13] T. Göbel, F. Breitinger, H. Baier, Optimising data set creation in the cybersecurity landscape with a special focus on digital forensics: Principles, characteristics, and use cases, Forensic Science International: Digital Investigation 52 (2025) 301882. doi:https://doi.org/10.1016/j.fsidi.2025.301882.

[14] S. Mombelli, J. R. Lyle, F. Breitinger, Fairness in digital forensics datasets' metadata–and how to improve it, Forensic Science International: Digital Investigation 48 (2024) 301681.

[15] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of management information systems 12 (1996) 5–33. 23 out. de 2023.

[16] W. W. W. C. (W3C), SHACL - shapes constraint language, https://www.w3.org/TR/shacl/, 2017. W3C Recommendation, 20 July 2017. Accessed June 2025.

[17] G. Klir, B. Yuan, Fuzzy sets and fuzzy logic, volume 4, Prentice hall New Jersey, 1995.

[18] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, et al., Fair principles: interpretations and implementation considerations, 2020.

[19] M. Ramzan, M. Shoaib, A. Altaf, S. Arshad, F. Iqbal, Á. K. Castilla, I. Ashraf, Distributed denial of service attack detection in network traffic using deep learning algorithm, Sensors 23 (2023) 8642.

[20] L. O. B. da Silva Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson, Fair data point: A fair-oriented approach for metadata publication, Data Intelligence 5 (2023) 163–183.

Uso de Inteligência Artificial Generativa e Ontologias para Indexação Temática de Imagens: Uma Abordagem Iconográfica de Panofsky

Adriana Aparecida Lemos Torres[1,*], *Alexandre Alves da* Rocha[1,2], *Benildes Coura Moreira dos Santos* Maculan[1], *Gislene Rodrigues da* Silva[1,5], *Felipe Moreira de* Assunção[3,4], *Francis Bento* Marques[1,6], *Elisângela Cristina* Aganette[1] *and Amanda Jercika Carla de Oliveira* Souza[1]

[1] *Universidade Federal de Minas Gerais, Programa de Pós-Graduação em Gestão e Organização do Conhecimento, Belo Horizonte/MG, Brasil*

[2] *Universidade Federal de Minas Gerais, Programa de Pós-graduação em Engenharia Elétrica, Belo Horizonte/MG, Brasil*

[3] *Universidade Federal de Minas Gerais, Laboratório Multiusuário de Computação Científica, Belo Horizonte/MG, Brasil*

[4] *Universidade Federal do Rio Grande do Sul, Laboratório de Dados, Métricas Institucionais e Reprodutibilidade Científica, Porto Alegre/RS, Brasil*

[5] *Universidade Carlos III de Madrid, Getafe/Madrid, Espanha*

[6] *Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina/MG, Brasil*

Resumo

Este artigo apresenta uma proposta metodológica para analisar de que forma modelos de inteligência artificial generativa podem ser empregados na representação de imagens de valor histórico e cultural. A pesquisa articula três elementos centrais: o método iconográfico de Erwin Panofsky, ontologias estruturadas aplicadas à organização do conhecimento e a avaliação dos resultados gerados pela inteligência artificial com base em métricas de similaridade semântica e concordância entre termos. Os experimentos realizados com diferentes versões de modelos de linguagem, como ChatGPT e Gemini, demonstram que, embora os sistemas de inteligência artificial sejam capazes de produzir descrições com sentido próximo às elaboradas por especialistas humanos, ainda apresentam dificuldades em adotar a mesma terminologia técnica utilizada por profissionais da área. Esse resultado reforça a relevância da curadoria humana nos processos de representação e indexação, sobretudo em contextos que demandam precisão conceitual. A proposta destaca o uso articulado de ferramentas tecnológicas e conhecimento especializado, apontando caminhos para uma aplicação mais responsável e eficiente da inteligência artificial na organização de acervos visuais.

Palavras-chave

Organização do conhecimento, ontologias, representação temática de imagens, inteligência artificial generativa, método iconográfico, curadoria humana

Abstract

This article presents a methodological proposal to analyze how generative artificial intelligence models can be used to describe images of historical and cultural value. The research combines three core elements: Erwin Panofsky's iconographic method, structured ontologies used for knowledge organization, and the evaluation of results generated by generative artificial intelligence based on metrics of meaning similarity and agreement between terms. Experiments conducted with different versions of language models, such as ChatGPT and Gemini, show that, although artificial intelligence systems are capable of producing descriptions with meaning close to those created by human experts, they still struggle to adopt the same technical terms used by professionals in the field. This reinforces the importance of human curation in description and indexing processes, especially in contexts that require conceptual precision. The proposal emphasizes the combined use of technological tools and specialized knowledge, indicating paths for a more responsible and efficient application of artificial intelligence in the organization of visual collections.

[1] *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.*

[*] Autor correspondente.

✉ adrianalemos.ufmg@gmail.com (A. A. L. Torres); alexandrear@ufmg.br (A. A. Rocha); benildes@gmail.com (B. C. M. S. Maculan); gislenerds@gmail.com (G. R. Silva); felipemoreira@ufmg.br (F. M. Assunção); fbmarques@gmail.com (F. B. Marques); elisangelaaganette@gmail.com (E. C. Aganette); amanda.jercika@gmail.com (A. J. C. Oliveira);

iD 0000-0002-3470-0676 (A. A. L. Torres); 0000-0003-4463-0162 (A. A. Rocha); 0000-0003-4303-9071 (B. C. M. S. Maculan); 0000-0003-2824-953X (G. R. Silva); 0000-0001-9975-5347 (F. M. Assunção); 0000-0001-8387-5272 (F. B. Marques); 0000-0003-4357-8016 (E. C. Aganette); 0009-0004-1026-0065 (A. J. C. Oliveira);

Keywords
Knowledge organization, ontologies, thematic representation of images, generative artificial intelligence, iconographic method, human curation

1. Introdução

A representação temática de imagens constitui um dos desafios mais complexos para a Ciência da Informação (CI), sobretudo quando se trata de representar acervos de relevância histórico-cultural. Fotografias de pinturas e esculturas, amplamente disseminadas em repositórios digitais, exemplificam documentos iconográficos que demandam múltiplos níveis de interpretação: desde a descrição de elementos visuais básicos até a compreensão de símbolos, contextos sociais e valores culturais. Tradicionalmente, o método iconográfico de Erwin Panofsky tem sido referência para tal tarefa, ao estruturar a análise em três níveis progressivos: pré-iconográfico, iconográfico e iconológico, permitindo capturar camadas de significado que extrapolam a superfície visível.

No entanto, a crescente escala dos acervos digitais e a diversidade de contextos de uso tornam a indexação manual insuficiente para atender demandas de volume, precisão, consistência e recuperação ampliada. Nesse cenário, tecnologias baseadas em Inteligência Artificial (IA), especialmente Modelos Largos de Linguagem (LLM) com capacidade generativa, surgem como alternativas para automatizar parte do processo de descrição. Ainda assim, a IA, isoladamente, carece de arcabouço interpretativo para alcançar profundidade iconográfica e iconológica de forma fidedigna.

Para preencher essa lacuna, este estudo propõe integrar ontologias do contexto histórico-cultural, o *Conceptual Reference Model of the International Committee for Documentation of the International Council of Museums* (CIDOC CRM); o Sistema de Classificação Iconográfica criado por Henri van de Waal para organizar e descrever imagens de obras de arte e outros materiais visuais (Iconclass) (vocabulário iconográfico estruturado em SKOS) e o *Art & Architecture Thesaurus* (Getty Research Institute) (Getty AAT) (tesauro em SKOS/ISO 25964), em complemento à ontologia CIDOC CRM. Esse conjunto fornece suporte semântico capaz de ampliar a consistência e a precisão conceitual das descrições geradas por IA. Assim, o objetivo central deste trabalho é avaliar a capacidade da inteligência artificial na representação temática de imagens, utilizando o método iconográfico de Erwin Panofsky, articulado ao uso de ontologias histórico-culturais para o enriquecimento semântico da representação.

Diante desse contexto, a questão de pesquisa que orienta este estudo é: em que medida a combinação de Inteligência Artificial Generativa (IAG) e ontologias do contexto histórico-cultural pode enriquecer a representação temática de imagens com base no método iconográfico de Erwin Panofsky, considerando as limitações da IAG na interpretação de significados simbólicos e a necessidade de curadoria humana para garantir a fidedignidade semântica? Assim, propõe-se que apenas a combinação entre IAG, ontologias estruturadas e curadoria humana seja capaz de garantir fidedignidade semântica na representação temática de imagens complexas

2. Bases teóricas

Esta seção apresenta os fundamentos teóricos que alicerçam esta pesquisa e trata sobre a representação temática de imagens, a utilização do método iconográfico de Erwin Panofsky e da inteligência artificial assim como o uso de ontologias do contexto histórico-cultural para esta indexação.

2.1. Representação temática de imagens

Na literatura da CI foi observado que não há consenso conceitual entre representação temática e indexação, ainda que indícios de diferenciação tenham sido apresentados por autores como Fujita

(1999), Guimarães (2003, 2008, 2009) e Lancaster (2004). A representação temática foi compreendida como atividade interpretativa de identificação dos temas centrais de um documento, enquanto a indexação foi definida como etapa descritiva, fundamentada nos resultados dessa interpretação, em que termos e categorias são atribuídos a partir de vocabulários controlados, tesauros ou ontologias.

A representação temática de imagens foi caracterizada como processo analítico que busca identificar, interpretar e expressar os temas centrais de recursos visuais, distinguindo-se da indexação por corresponder a uma fase cognitiva anterior, voltada à leitura e abstração de significados. A indexação, em contrapartida, foi descrita como atribuição formal de termos ou códigos em sistemas documentários, com vistas à organização e recuperação da informação (MAIMONE; GRACIOSO, 2007).

Foi ressaltado que a indexação de imagens permanece como desafio, devido à complexidade dos atributos visuais e à diversidade de interpretações possíveis. Nesse cenário, a fotografia foi destacada como tipologia documental dotada de linguagem própria, assumindo múltiplas funções em contextos de informação, memória, arte e comunicação (TORRES; MACULAN, 2023; TORRES et al., 2025).

Como síntese, foi enfatizado que a representação temática constitui um macroprocesso dividido em análise conceitual (interpretação e identificação de temas e relações) e tradução (seleção e normalização em vocabulários). A indexação correspondeu a esse segundo momento, no qual os descritores materializam a representação nos sistemas, evitando sobreposições terminológicas e permitindo a avaliação de resultados de Inteligência Artificial em estudos da área.

2.2. Método Iconográfico de Erwin Panofsky

A necessidade de mitigar a subjetividade dos indexadores e de aprimorar a eficiência da indexação foi apontada como razão para a adoção de metodologias específicas na representação de documentos iconográficos, devido às diferenças em relação aos documentos textuais (MANINI, 2002; SMIT, 1996; TORRES, 2019). Entre os modelos discutidos na literatura, destacou-se o método iconográfico ou iconológico de Erwin Panofsky (1986), originalmente desenvolvido no campo da História da Arte e posteriormente incorporado em análises no âmbito da CI. Esse modelo foi reconhecido por sua contribuição à compreensão tanto de elementos visuais quanto de significados intrínsecos e extrínsecos, em articulação com o contexto histórico-cultural de origem.

A metodologia foi estruturada em três níveis: o pré-iconográfico, voltado à identificação de elementos formais básicos; o iconográfico, dedicado ao reconhecimento de temas e motivos mediante conhecimento cultural; e o iconológico, destinado à interpretação de significados simbólicos e contextuais mais profundos, fundamentados em pesquisa histórica, social e cultural (SILVA; DIAS, 2018). Apesar das críticas relativas à sua rigidez metodológica (GOMBRICH, 1986), o método permanece como referência para a indexação semântica de acervos histórico-culturais e para a área da CI.2.3

2.3. Inteligência Artificial Generativa

A presença da IA tem sido ampliada em diversos setores da sociedade, sendo observado que atividades relacionadas à representação da informação também passaram a ser impactadas. Nesse contexto, a IAG foi destacada como vertente em rápida expansão, ao possibilitar a produção de textos e imagens por sistemas computacionais, redefinindo formas de análise e representação de objetos visuais. Desde o surgimento, no final de 2022, dos primeiros modelos de IAG baseados em *Large Language Models* (LLM), sua aplicação em diferentes áreas foi constatada, com ênfase para a CI, em função do interesse pela forma como o conhecimento é representado nesses sistemas. Ressaltou-se, ainda, a importância dos elementos semânticos estruturados nos Sistemas de Organização do Conhecimento (SOC) para a representação temática (SILVA; DIAS, 2024).

Foi evidenciado que os LLM se configuram como tecnologias proeminentes por sua capacidade de processar e gerar linguagem natural e imagens, podendo apoiar significativamente a representação de fotografias e sua descrição temática. Tais modelos, treinados com grandes volumes de dados e baseados em redes neurais profundas, operam por inferências estatísticas para gerar textos contextualizados (BOMMASANI et al., 2021). Em contextos culturais, entretanto, foi ressaltado que os LLM não são neutros, mas agentes que também modelam narrativas e influenciam representações simbólicas (BENDER et al., 2021).

Modelos como o ChatGPT, da OpenAI, e o Gemini, do Google, foram reconhecidos como recursos promissores na mediação informacional, ao integrarem visão computacional e processamento de linguagem natural para o reconhecimento de padrões visuais e a inferência de significados. Foi apontado que avanços relevantes vêm sendo obtidos na qualidade da representação de dados e metadados, com reflexos positivos na catalogação de acervos digitais (ZENG, 2019; MARTINS et al., 2022). Além disso, quando utilizados em conjunto com ontologias, os LLM permitem mitigar problemas de heterogeneidade semântica, ambiguidade conceitual e ausência de contexto, viabilizando descrições e vínculos informacionais fundamentados em conceitos, instâncias, propriedades e relações interpretáveis tanto por humanos quanto por máquinas (FARINELLI; SOUZA, 2021).

2.4. Sistemas de Organização do Conhecimento no patrimônio cultural

Ontologias como a CIDOC CRM, reconhecido como padrão internacional ISO 21127:2006, desenvolvido pelo ICOM (International Council of Museums), e vocabulários controlados como Getty AAT e Iconclass são reconhecidas internacionalmente por modelar conhecimento em museus, bibliotecas e arquivos. Sua integração com IA visa garantir precisão, interoperabilidade e consistência terminológica (DOERR, 2003). Tais instrumentos de modelagem conceitual vêm sendo adotados na Ciência da Informação para a representação descritiva, a organização do conhecimento e a interoperabilidade entre sistemas, ao permitir a formalização de entidades, propriedades e relações (SILVA; SOUZA, 2014). Foi observado que esses modelos podem superar limitações de sistemas baseados apenas em palavras-chave, ao possibilitar recuperação contextualizada (RAMALHO, 2010) e enriquecer a representação descritiva de documentos em ambientes digitais (OLIVEIRA, 2020).

Destaca-se que a atribuição de semântica à informação é realizada por meio de metadados, que descrevem elementos em documentos textuais e características técnicas e semânticas em documentos multimídia (GILLILAND-SWETLAND, 2000; SILVA; SOUZA, 2014). No campo do patrimônio cultural, o CIDOC CRM foi descrito como ontologia de referência, projetada para representar entidades, eventos, objetos e relações, possibilitando a contextualização de objetos culturais e a realização de inferências a partir de dados complexos (SILVA; SOUZA, 2014). De forma complementar, vocabulários controlados como Getty AAT e Iconclass foram utilizados para descrever conceitos artísticos e iconográficos, assegurando granularidade e padronização terminológica. Também foi destacada a utilização da Wikidata como ambiente de interligação semântica e publicação de dados culturais em acesso aberto (MARCONDES, 2016).

No contexto europeu, o Europeana Data Model (EDM) foi desenvolvido como estrutura semântica para promover interoperabilidade entre instituições culturais, fundamentado em padrões da Web Semântica como RDF(S), SKOS, OAI-ORE e Dublin Core (DOERR et al., 2010). Esse modelo adota abordagens centradas em objetos e eventos, permitindo associações mais ricas e maior contextualização de dados (CARRASCO; VIDOTTI, 2018), além de viabilizar a publicação de metadados como *Linked Open Data*, conectando informações culturais a outros recursos na Web (HASLHOFER; ISAAC, 2011).

3. Metodologia

Este estudo adota uma abordagem exploratória, pois visa compreender um campo ainda recente na CI, que é o uso IAG no processo de indexação de imagens utilizando um método já consolidado, que é o proposto por Panofsky. Essa aplicação foi articulada com o uso de ontologias e vocabulários controlados. A pesquisa é considerada de natureza qualitativa, pois tem o objetivo de compreender o fenômeno do uso da IAG no processo de indexação, que até então, era realizada majoritariamente por humanos. Também se aplica a análise quantitativa ao comparar a capacidade da IAG elaborar termos relevantes em relação ao Modelo Referencial Colaborativo de Indexação (MRCI). Dessa forma, aplicou-se métricas estatísticas, como o coeficiente Kappa de Cohen e a similaridade de cosseno baseada em *embeddings* semânticos. A partir disso, comparou-se a proximidade entre a indexação realizada por humanos e as que foram geradas por IAG.

3.1. Justificativa da abordagem

A escolha do método iconográfico de Panofsky justifica-se por sua robustez conceitual e sua tradição consolidada tanto na História da Arte quanto na CI. Diferentemente de abordagens baseadas apenas em descritores formais ou metadados técnicos, Panofsky organiza a leitura em três níveis complementares, ideal para testar em que medida a IAG consegue avançar da leitura factual automatizada para níveis de interpretação mais profundos, mediada por ontologias e curadoria humana.

A IAG, representada pelo ChatGPT e pelo *Gemini*, foi escolhida por sua comprovada capacidade de gerar texto coeso a partir de insumos visuais, por sua base multimodal robusta e por sua acessibilidade para aplicações acadêmicas. Entretanto, reconhece-se que modelos generalistas não foram treinados especificamente para interpretação iconográfica, o que torna este teste pioneiro em demonstrar potenciais e lacunas.

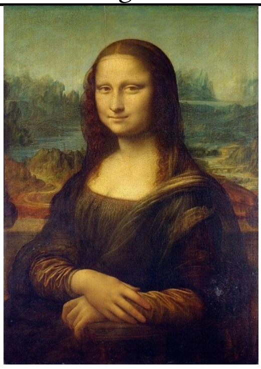
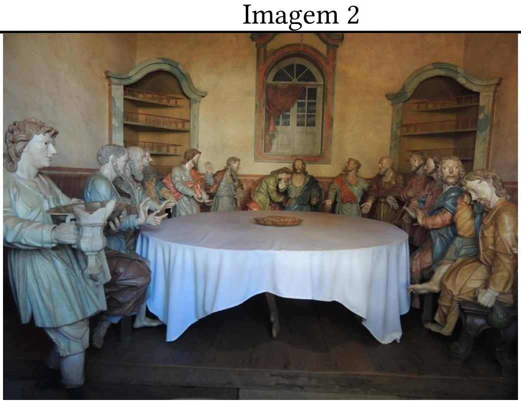
Para enriquecer semanticamente a descrição, as saídas geradas pela IAG serão mapeadas e refinadas por meio de SOC reconhecidos no campo do patrimônio cultural: a ontologia CIDOC CRM e os vocabulários controlados (em SKOS) Iconclass, Getty AAT, Wikidata e EDM. Esta seleção baseia-se em referencial teórico que aponta sua robustez em capturar relações conceituais, hierarquias e vocabulário controlado no contexto histórico-cultural, fundamentais para representar as imagens deste contexto e nos níveis iconográfico e iconológico do modelo de Panofsky. Nesse contexto, as ontologias e os vocabulários controlados histórico-culturais selecionados cumprem o papel de suporte semântico estruturado, oferecendo controle de vocabulário, relações conceituais e contexto histórico necessários para enriquecer a descrição automática e reduzir ambiguidades.

3.2. Amostra de imagens

O corpus foi composto por imagens fotográficas de duas tipologias distintas de obras de arte: uma pintura (Mona Lisa) e um conjunto de esculturas em madeira e pedra do Aleijadinho (A Última Ceia), conforme apresentado no quadro 1.

Quadro 1

Amostra de imagens usadas para a indexação

Imagem 1	Imagem 2
	
Pintura: Mona Lisa (1503-1506) Pintor: Leonardo da Vinci Fotografia: C2RMF: Galerie de tableaux en très haute définition	Esculturas: A Última Ceia Escultor: Antônio Francisco Lisboa, o Aleijadinho (1730- 1814) Fotógrafa: Lila Cruz

Fonte:Wikimedia Commons - https://w.wiki/6C2o	Fonte: Pixabay - https://pixabay.com/pt/photos/jesus-escultura-viajar-por-am%C3%A9rica-2291128/

Fonte: elaborado pelos autores (2025).

A seleção destas obras se fundamenta em quatro critérios principais: 1) Contexto comum: cenário histórico-cultural; 2) Consagração cultural: são objetos de estudo amplamente referenciados, o que facilita a validação das interpretações automatizadas com descrições especializadas já consolidadas; 3) Contraste técnico e material: pintura e escultura apresentam diferenças na técnica de execução, materialidade, cor, forma e textura, o que permite analisar se tais variações impactam o desempenho da IAG; e 4) Unificação documental: ambas as obras são representadas no experimento por fotografias digitais, igualando o formato de entrada para os modelos de IA. Além disso, se baseia em Panofsky (1986) quando destaca que o uso de obras canônicas favorece a validação, pois apresentam interpretações iconográficas consolidadas, o que permite contrastar os resultados com leituras já reconhecidas.

3.3. Testes

Foram conduzidos dois testes distintos para cada imagem analisada. O Teste 1 (T1) consistiu na aplicação do método de Panofsky a partir de um único *prompt*, no qual se solicitava a geração de três listas correspondentes aos níveis pré-iconográfico, iconográfico e iconológico. As instruções determinavam o uso de termos e códigos de vocabulários controlados (Iconclass, Getty AAT, CIDOC CRM), com a exigência de indicar "sem correspondência" quando não houvesse alinhamento direto. O Teste 2 (T2) adotou uma estratégia encadeada de *prompts*. Nesse caso, o modelo recebia instruções sucessivas: (i) listar apenas elementos visuais observáveis (nível pré-iconográfico), (ii) identificar temas/motivos (nível iconográfico) e (iii) interpretar significados e valores culturais (nível iconológico). Após essa sequência, aplicou-se um quarto *prompt* idêntico ao T1, de modo a gerar uma saída consolidada e permitir a comparação entre os dois desenhos experimentais. A engenharia de *prompts* é sintetizada no quadro 2.

Quadro 2

Esquema de Testes e Instruções segundo os Níveis de Panofsky

Teste	Exemplo de instrução	Objetivo
T1	"Leia a imagem X e produza três listas correspondentes aos níveis de Panofsky (pré-iconográfico, iconográfico e iconológico). Sempre que possível, utilize termos/códigos de Iconclass, AAT e CIDOC CRM; quando não houver correspondência, indique explicitamente 'sem correspondência'."	Avaliar desempenho em um único prompt
T2a	"Liste apenas os elementos visuais observáveis (formas, cores, personagens), mapeando-os para termos do AAT e Iconclass quando aplicável."	Pré-iconográfico
T2b	"Identifique os temas/motivos presentes na imagem, utilizando códigos do Iconclass."	Iconográfico
T2c	"Explique os significados e valores culturais subjacentes, registrando termos de AAT ou CIDOC CRM quando disponíveis."	Iconológico
T2d	(idêntico ao T1)	Consolidação para comparação

Fonte: elaborado pelos autores (2025).

Diversas iterações foram realizadas até a definição final dos *prompts*. Todos os experimentos foram conduzidos em ambiente anônimo, sem memória de conversas anteriores e com os parâmetros padrão dos modelos, sem calibragem adicional de *temperature*, *top-p* ou limite de *tokens*. Essa padronização assegura que as diferenças nos resultados decorrem do comportamento intrínseco dos modelos, e não de ajustes externos.

Importante destacar que os modelos de IAG não apresentam comportamento determinístico: a geração textual é um processo probabilístico, resultando em múltiplas saídas possíveis para o mesmo *input* (HOLTZMAN et al., 2019; RADFORD et al., 2019; ZHANG et al., 2020). Para mitigar esse efeito, foi executada uma rodada padronizada de testes, assegurando consistência e comparabilidade entre T1 e T2.

3.4. Procedimentos

O experimento segue cinco etapas articuladas: 1) Seleção e padronização do corpus de imagens, garantindo qualidade técnica e contexto histórico homogêneo; 2) Anotação manual por especialistas com base no método de Panofsky, categorizando cada imagem nos três níveis (pré-iconográfico, iconográfico e iconológico); 3) Apresentação do MRCI. O MRCI foi construído por 5 especialistas da área de Biblioteconomia e CI, que realizaram anotações independentes das imagens segundo os três níveis de Panofsky, utilizando como referência os vocabulários controlados (AAT, Iconclass) e a ontologia CIDOC CRM. Em seguida, aplicou-se um processo de consenso: discrepâncias foram discutidas coletivamente, registrando-se as decisões de inclusão ou exclusão de termos, com preferência por rótulos prefLabel em SKOS. O resultado final foi consolidado em listas estruturadas por nível, contendo rótulos e códigos correspondentes sempre que disponíveis. O MRCI constitui o padrão humano de referência para comparação com as saídas das IAGs; 4) Geração de descrições automáticas utilizando ChatGPT e Gemini, com prompts ajustados para leitura nos três níveis de Panofsky e com a orientação de recorrer aos SOC selecionados (CIDOC CRM, AAT, Iconclass, EDM, Wikidata); 5) Validação e comparação dos resultados. As saídas das IAG foram verificadas quanto a inconsistências, lacunas e imprecisões, sendo confrontadas com o MRCI. Essa comparação permitiu mensurar o grau de alinhamento entre indexação humana e automática.

3.5. Análise de resultados: Métricas qualitativas e quantitativas

A análise integrará dimensões qualitativas e quantitativas. No aspecto qualitativo, serão avaliadas coerência, profundidade interpretativa e consistência entre os níveis de Panofsky nas descrições geradas e conformidade com as ontologias utilizadas.

No aspecto quantitativo, a avaliação será baseada em métricas estatísticas utilizadas para mensurar a concordância entre classificadores, incluindo o coeficiente Kappa de Cohen e a similaridade de cosseno (COHEN, 1960). O coeficiente Kappa será aplicado para quantificar a concordância entre as categorias atribuídas pelo Modelo Referencial Colaborativo de Indexação (MRCI) e pelas IA, considerando a influência de concordâncias ocorridas ao acaso. Para isso, as descrições serão organizadas em matriz de contingência, a partir da qual serão calculadas as proporções de concordância observada (P_0) e esperada ao acaso (P_e), segundo a equação $k = (P_0 - P_e) / (1 - P_e)$, em que valores próximos de 1 indicam alta concordância, e valores inferiores a 0,40, baixa correspondência.

A similaridade de cosseno será utilizada para examinar a proximidade semântica entre descritores humanos e aqueles gerados pelas IA. As descrições textuais serão convertidas em representações vetoriais por meio do modelo BERTimbau, que captura relações semânticas entre palavras e expressões (SOUZA; NOGUEIRA; LOTUFO, 2020). A equação $\cos(\theta) = (A \cdot B) / (|A| \cdot |B|)$ permitirá avaliar se a IA interpreta corretamente não apenas os elementos descritivos explícitos, mas também os significados subjacentes nas análises iconográficas, sendo valores próximos de 1 indicativos de alta semelhança semântica, e valores inferiores a 0,5, de baixa correlação.

4. Resultados e análises

A avaliação da indexação foi conduzida por meio de uma abordagem comparativa entre a atividade realizada por especialistas humanos e a gerada por inteligência artificial. O objetivo principal desta análise foi determinar a eficiência e as limitações dos modelos de IA empregados, verificando sua capacidade de replicar estruturas interpretativas estabelecidas por metodologias tradicionais e com o uso de ontologias e tesauros.

Os principais aspectos considerados na avaliação incluem: 1) Precisão da indexação: Grau de concordância entre os descritores atribuídos pela IA e aqueles definidos por especialistas; 2) Capacidade de abstração: Identificação de padrões semânticos e contextuais em diferentes níveis de interpretação iconográfica; 3) Consistência na categorização: Uniformidade dos descritores gerados ao longo do conjunto de dados;4) Identificação de limitações: Diagnóstico de possíveis falhas do modelo, como dificuldades na interpretação de simbolismos culturais ou presença de viés algorítmico.

4.1. Resultados da indexação

Após a realização das indexações das duas imagens da amostra - Mona Lisa e A Última Ceia - pelas ferramentas de IAG (ChatGPT e *Gemini*) utilizando dois testes (T1 e T2), os resultados obtidos foram tratados, tendo alguns exemplos expostos no quadro 2.

Os testes foram realizados com duas versões do ChatGPT (ChatGPT 4o e ChatGPT 4o-mini) e duas versões do *Gemini* (*Gemini Flash 2.5 e Gemini Flash Pro*). O quadro 3 apresenta apenas uma seleção dos resultados: o teste T1 realizado com o Gemini Flash Pro, o teste T2 com o ChatGPT, além do MRCI. De modo geral, observou-se que os resultados variaram entre as versões, e que cada uma também produziu respostas distintas nos dois tipos de teste (T1 e T2). No primeiro, foi utilizado um único prompt; no segundo, adotou-se uma sequência de prompts interdependentes, com o objetivo de fornecer maior contexto à IAG. Após essa construção de contexto, aplicou-se o mesmo prompt final utilizado no T1. O MRCI apresentado, foi o padrão utilizado para as análises quantitativas apresentadas na próxima seção.

Quadro 3
Resultados das indexações da imagem da Mona Lisa realizadas pelas IAG e por profissionais da CI

IA ou Profissional	1 - Pré-iconográfico	2 - Iconográfico	3 - Iconológico
Gemini Flash Pro (T1)	Dress, Landscape, Man-Made Object, Oil paint, Painting (visual work), Portrait of a woman, Sitting, Woman	Creation, Leonardo da Vinci, Lisa Gherardini, Louvre Museum, Mona Lisa, Portrait, Portrait of historical personage, 'LISA DEL GIOCONDO'	Aerial perspective, Beauty, 'Venustas'; 'Bellezza' (Ripa), Propositional Object, High Renaissance, Humanism, Idealism, providedCHO (Provided Cultural Heritage Object), Sfumato
ChatGPT 4o (T2)	Oil painting, Panel painting, Physical Man-Made Stuff, Physical Thing, Portraits of individuals	Lisa del Giocondo, Mona Lisa, Portrait, Portrait of a woman, Renaissance	Actor, Time-Span, Place, Recurso Cultural Provido, Sfumato, Mystery, Landscape
Profissional humano (Modelo Referencial Colaborativo de Indexação)	Background Landscape, Dark Dress, Frontal View, Half-Length, Hands One Over the Other, Painting, Portrait, Seated Woman, Soft Lighting (Sfumato), Woman	15th-16th Century Painting Technique, Enigmatic Smile, Harmony Between Figure and Background Nature, Leonardo da Vinci, Lisa Gherardini, Louvre Museum, Mona Lisa, Oil Painting on Wood, Renaissance, Renaissance Portrait	Art as Expression of the Individual, Enigmatic Expression, Humanism, Icon of Western Art, Ideal Beauty, Realism, Renaissance, Renaissance Art, Renaissance Ideal of Female Beauty, Sfumato Technique

Fonte: elaborado pelos autores (2025).

4.2. Análises de coeficientes

A aplicação do coeficiente Kappa de Cohen e da medida de similaridade semântica obtida a partir dos vetores gerados pelo modelo BERTimbau permitiu avaliar, com critérios complementares, a correspondência entre os resultados produzidos pelas IAG e o Modelo Referencial Colaborativo de

Indexação, considerando tanto a concordância terminológica quanto a proximidade de sentido entre os termos utilizados.

Os resultados de similaridades das indexações das IAG referente à imagem da Mona Lisa frente ao MRCI são apresentados no quadro 4 e mostram alta similaridade semântica contrapondo à baixa similaridade de termos (leve ou pobre).

Quadro 4
Resultados de similaridades das indexações da imagem da Mona Lisa realizadas pelas IAG e por profissionais da CI.

Avaliador1	Avaliador 2 (Profissionais da CI)	Similaridade BERTimbau	Classificação BERTimbau	Coeficiente Kappa	Classificação Kappa
ChatGPT 4º (T1)	MRCI	**0.81**	Alta Similaridade	0.00	Leve
ChatGPT 4º (T2)	MRCI	0.93	Alta Similaridade	0.00	Leve
ChatGPT 4o-mini (T1)	MRCI	0.87	Alta Similaridade	0.00	Leve
ChatGPT 4o-mini (T2)	MRCI	**0.95**	Alta Similaridade	0.00	Leve
Gemini Flash 2.5 (T1)	MRCI	0.92	Alta Similaridade	0.00	Leve
Gemini Flash 2.5 (T2)	MRCI	0.90	Alta Similaridade	0.00	Leve
Gemini Flash Pro (T1)	MRCI	0.92	Alta Similaridade	**0.10**	**Leve**
Gemini Flash Pro (T2)	MRCI	0.88	Alta Similaridade	**-0.10**	**Pobre**

Fonte: elaborado pelos autores (2025).

Os dados revelam um padrão recorrente: embora todos os modelos de IAG tenham alcançado altos índices de similaridade semântica — com valores variando de 0.81 a 0.95 nas métricas obtidas via modelo BERTimbau —, os valores do coeficiente Kappa de Cohen permaneceram consistentemente baixos, oscilando entre 0.00 e -0.10, indicando leve ou pobre concordância em termos exatos utilizados. Esse contraste evidencia que, apesar de os modelos conseguirem captar o sentido geral ou semântico das indexações humanas, eles não reproduzem com precisão os mesmos termos escolhidos pelos profissionais que elaboraram o MRCI. Ou seja, há convergência de sentido, mas divergência na escolha lexical ou terminológica, o que é particularmente relevante em contextos de representação do conhecimento e organização da informação, onde a terminologia controlada é fundamental.

Destaca-se que os testes T2, que envolveram a aplicação de prompts em sequência para construção de contexto, não resultaram em ganhos na concordância terminológica em comparação aos testes do tipo T1, que utilizaram apenas um prompt. Isso indica que a introdução de mais contexto semântico não foi suficiente para alinhar a linguagem das IAGs à terminologia especializada usada por indexadores humanos. Observa-se, ainda, que entre os modelos avaliados, o *Gemini Flash Pro (T2)* foi o único a apresentar um coeficiente Kappa negativo (-0.10), classificado como "pobre", sugerindo uma tendência à discordância terminológica mesmo quando há entendimento semântico global.

O quadro 5 apresenta os resultados de similaridades das indexações das IAG referente à imagem de A Última Ceia frente ao MRCI. Da mesma maneira que ocorreu com os resultados da Mona Lisa, os resultados mostraram alta similaridade semântica contrapondo à baixa similaridade de termos (leve ou pobre).

Quadro 5
Resultados de similaridades das indexações da imagem de A Última Ceia realizadas pelas IAG e por profissionais da CI

Avaliador1	Avaliador 2 (Profissionais da CI)	Similaridade BERTimbau	Classificação BERTimbau	Coeficiente Kappa	Classificação Kappa
ChatGPT 4° (T1)	MRCI	**0.93**	Alta Similaridade	0.00	Leve
ChatGPT 4° (T2)	MRCI	**0.84**	Alta Similaridade	0.00	Leve
ChatGPT 4o-mini (T1)	MRCI	0.91	Alta Similaridade	0.00	Leve
ChatGPT 4o-mini (T2)	MRCI	**0.93**	Alta Similaridade	**0.11**	Leve
Gemini Flash 2.5 (T1)	MRCI	0.86	Alta Similaridade	-0.05	Pobre
Gemini Flash 2.5 (T2)	MRCI	0.88	Alta Similaridade	0.00	Leve
Gemini Flash Pro (T1)	MRCI	0.87	Alta Similaridade	**-0.11**	Pobre
Gemini Flash Pro (T2)	MRCI	0.88	Alta Similaridade	0.00	Leve

Fonte: elaborado pelos autores (2025).

Os resultados obtidos para a imagem A Última Ceia mantêm o padrão identificado anteriormente, mas adicionam nuances importantes à interpretação geral sobre o desempenho das IAG em tarefas de indexação. Embora os índices de similaridade semântica permaneçam elevados em todas as combinações de modelo e tipo de teste, observa-se uma ligeira ampliação na variação dos coeficientes Kappa, com destaque para dois casos de concordância classificada como pobre, ambos registrados nos teste T1 dos modelos *Gemini Flash 2.5* e *Gemini Flash Pro.* A presença desses coeficientes negativos sugere não apenas uma baixa sobreposição de termos, mas também uma tendência à divergência sistemática em relação à terminologia adotada pelos profissionais da área. Isso indica que, mesmo quando os modelos compreendem corretamente o conteúdo ou sentido da imagem — como evidenciado pelos altos valores de similaridade semântica —, as escolhas lexicais feitas pelas IAGs podem se afastar significativamente dos critérios técnicos e padronizados de indexação humana.

Outro aspecto que merece atenção é a sutil diferença nos desempenhos entre os testes T1 e T2, especialmente no caso do modelo ChatGPT 4o-mini, que apresentou leve aumento na concordância terminológica no teste T2. Esse comportamento, ainda que limitado, sugere que a introdução de contexto adicional pode exercer influência positiva, ainda que modesta, na capacidade das IAGs de se aproximarem das escolhas humanas em termos terminológicos. No entanto, essa variação pontual não foi suficiente para alterar a tendência geral, e permanece o indicativo de que a estratégia de prompts interdependentes, tal como aplicada neste estudo, não constituiu uma solução determinante para a convergência terminológica.

Essa constatação abre espaço para reflexões metodológicas mais amplas. A manutenção de alta similaridade semântica associada a baixa concordância de termos reforça a ideia de que as IAGs operam com modelo de linguagem baseado em probabilidade contextual, e não em sistemas conceituais formalizados, como é o caso de tesauros, ontologias ou vocabulários controlados utilizados por profissionais da área de representação do conhecimento. Tal descompasso entre compreensão de sentido e aderência terminológica evidencia a necessidade de estratégias complementares, como o alinhamento das saídas das IAGs às bases terminológicas estruturadas ou a incorporação de mecanismos de restrição lexical nas interações com os modelos.

Portanto, os resultados referentes à imagem A Última Ceia não apenas reforçam os achados observados na análise da Mona Lisa, mas também introduzem elementos que tornam ainda mais evidente o desafio da integração plena entre sistemas baseados em inteligência artificial e práticas tradicionais de representação do conhecimento. O conjunto dos dados analisados sugere que, no estado atual da tecnologia, as IAGs se mostram mais adequadas como instrumentos de apoio à geração de descritores preliminares, cabendo aos profissionais a função de validar, corrigir e alinhar esses termos às estruturas conceituais e linguísticas de vocabulários controlados e ontologias.

4.3. Análises de resultados

A partir das análises realizadas neste estudo, ficou evidente que os modelos de IAG, embora apresentem elevada competência na formulação de descrições semanticamente coerentes, ainda enfrentam dificuldades substanciais no alinhamento terminológico com vocabulários controlados. Este descompasso não apenas reforça os limites técnicos desses sistemas, mas também se insere em uma discussão mais ampla sobre a confiabilidade da informação gerada por IAs, especialmente à luz do fenômeno conhecido como alucinação.

As alucinações em IAG, conforme discutido por Barria Huidobro (2024), são entendidas como a geração de conteúdos imprecisos, enganosos ou factualmente incorretos, mesmo quando a estrutura sintática e semântica parece plausível. Embora o presente estudo não tenha se dedicado especificamente à detecção de alucinações, a ausência de concordância terminológica, mesmo diante de alta similaridade semântica, pode sinalizar traços iniciais desse fenômeno em domínios especializados como a indexação de imagens. Foi possível verificar nos resultados várias divergências entre as informações produzidas pelas IAGs e as informações reais das ontologias e tesauros analisados, como os descritores e seus respectivos códigos, descrições e fontes. As informações errôneas obtidas pelas IAG são consideradas graves, especialmente quando se trata de contextos que demandam precisão conceitual e terminológica e aderência a sistemas normativos de representação do conhecimento.

Estudos como o de Correa Busquets & Maccarini Llorens (2023) reforçam a importância de mecanismos de autossupervisão e modelos de verificação para mitigar tais ocorrências. No campo da representação da informação, isso pode significar a integração de mecanismos de validação semântica e terminológica diretamente nos fluxos de trabalho com IAG, combinando as potencialidades da IA com a robustez das estruturas ontológicas, tesauros e outras formas de controle de vocabulário já consolidadas. Essa perspectiva se alinha ao conceito de inteligência artificial aumentada, em que o conhecimento produzido pelas IAGs é avaliado, complementado e ajustado por profissionais humanos, promovendo um modelo colaborativo que valoriza tanto a escala e agilidade dos sistemas computacionais quanto o julgamento crítico e a expertise conceitual da curadoria humana (SOUZA; LIMA, 2025). Sob uma perspectiva mais ampla, Lemos (2024) propõe que falhas e erros nos sistemas digitais, incluindo as alucinações, podem funcionar como elementos reveladores de camadas profundas dos objetos e das mediações tecnológicas. Assim, o comportamento dos modelos de linguagem ao descrever obras visuais clássicas, como A Última Ceia e Mona Lisa, revela tensões epistemológicas entre o funcionamento algorítmico probabilístico e as exigências de precisão conceitual próprias da indexação humana.

Diante disso, os resultados apresentados neste estudo não apenas evidenciam a atual capacidade das IAG na compreensão de conteúdos complexos, mas também apontam para a necessidade de desenvolvimento de métodos híbridos, nos quais a colaboração entre máquinas e especialistas seja mediada por critérios técnicos claros, mecanismos de verificação e sensibilidade crítica.

5. Considerações finais

O estudo foi apresentado como contribuição ao campo das ontologias aplicadas, ao se explorar seu uso prático na qualificação de sistemas de Inteligência Artificial Generativa voltados à representação de documentos iconográficos. Foi evidenciado que a inovação metodológica se encontra na triangulação entre o método iconográfico de Panofsky, o emprego de ontologias controladas e a avaliação automatizada por coeficiente Kappa de Cohen e similaridade semântica baseada no modelo BERTimbau, organizados em um fluxo integrado de análise. A integração de indexação temática, ontologias e IA generativa em protocolo auditável foi destacada como abordagem inédita.

Foi ressaltada a relevância social da pesquisa, uma vez que o papel humano na validação e curadoria do conhecimento produzido foi considerado indispensável, demonstrando a complementaridade entre expertise humana e inteligência artificial. A análise de obras de Leonardo da Vinci e Aleijadinho foi utilizada como contraponto metodológico para valorizar semelhanças e diferenças no processo de indexação automática.

Os resultados mostraram que, embora alta similaridade semântica com referenciais humanos tenha sido alcançada pelas IAG, a concordância terminológica permaneceu baixa, revelando limitações quanto à aderência a esquemas conceituais formalizados. Reforçou-se, assim, a necessidade da curadoria humana para garantir precisão e consistência terminológica, especialmente em contextos que exigem padronização e interoperabilidade semântica.

Foi indicado que a pesquisa possui caráter exploratório, sendo recomendada sua continuidade por meio da ampliação do corpus, do refinamento das ontologias, da incorporação de *datasets* com metadados controlados e da aplicação dos métodos em acervos reais. Tais perspectivas foram apontadas para consolidação de abordagens híbridas entre inteligência artificial e conhecimento estruturado como caminhos eficazes, éticos e sustentáveis na preservação, interpretação e acesso ao patrimônio cultural.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio à pesquisa (Processo 307765/2023-7; Autor 3 e Processo 306386/2025-9, Autor 7). A Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) – Programa de Apoio à Pós-Graduação (PAG) pelo apoio à pesquisa (Identificador 13473, Autor 8).

Declaração sobre IA Generativa

Neste trabalho, foi explorado o uso de inteligência artificial generativa, especificamente os modelos ChatGPT (4o e o4-mini) e *Gemini (2.5 Flash e 2.5 Pro*), em conjunto com ontologias, para a análise e representação temática de imagens. As ferramentas foram empregadas na identificação de elementos visuais e no apoio à interpretação de significados em níveis pré-iconográfico, iconográfico e iconológico, com foco em tarefas que demandam conhecimento técnico especializado. Todo o processo foi acompanhado pela equipe de pesquisa, e os resultados foram revisados e validados, sendo assumida integral responsabilidade pelos conteúdos apresentados.

Referências

[1] C. Barria Huidobro. Alucinaciones de la inteligencia artificial: impacto en tecnología, política y sociedad. *Revista Estrategia, Poder y Desarrollo.* 2024.

[2] E. M. Bender, T. Gebru, A. Mcmillan-Major, S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? *In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21),* Canada, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.

[3] R. Bommasani et al. On the Opportunities and Risks of Foundation Models. *ArXiv*, 2021. URL: https://arxiv.org/abs/2108.07258.

[4] S. C. Busquets, L. M. Lorens. Self-supervision of Hallucinations in Large Language Models: LLteaM. *Journal of Language and Computation Research*, 2023. doi: 10.4995/jclr.2023.20408.

[5] L. B. Carrasco, S. Vidotti. *Patrimônio cultural: um panorama do modelo de dados da Europeana*. 2018.

[6] J. A. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, v. 20, n. 1, p. 37-46, 1960. doi: 10.1177/001316446002000104.

[7] M. Doerr. The CIDOC CRM — an ontological approach to semantic interoperability of metadata. *AI Magazine*, v. 24, n. 3, p. 75–92, 2003.

[8] M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, H. Van de Sompel. *The Europeana Data Model (EDM).* 2010. URL: https://www.researchgate.net/publication/303058300.

[9] F. Farinelli, A. Souza. *Ontologias de alto nível: porque precisamos e como usar.* 1. 2021, pp. 174-202.

[10] M. S. L. Fujita. A leitura do indexador: estudo de observação. *Perspectivas em Ciência da Informação*, Belo Horizonte, v. 4, n. 1, 1999, pp. 101–116. URL: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/597.

[11] A. J. Gilliland-Swetland. Enduring paradigm, new opportunities: the value of the archival perspective in the digital environment. Washington, DC: *Council on Library and Information Resources*, Digital Library Federation, 2000. URL: https://files.eric.ed.gov/fulltext/ED440660.pdf.

[12] E. H. Gombrich. *Arte e Ilusão: um estudo da psicologia da representação pictórica.* 5. ed. Rio de Janeiro: LTC, 1986.

[13] E. H. Gombrich. *Imágenes simbólicas: estudios sobre el arte del Renacimiento.* Madrid: Alianza, 1986.

[14] J. A. C. Guimarães. A análise documentária no âmbito do tratamento da informação: elementos históricos e conceituais. In: Rodrigues GM, Lopes IL. (org.). *Organização e representação do conhecimento na perspectiva da Ciência da Informação.* Brasília: Thesaurus, 2003, pp. 100–117.

[15] J. A. C. Guimarães. A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knowledge Organization (ISKO). *Revista Ibero-americana de Ciência da Informação.* Brasília, v. 1, n. 1, jan./abr. 2008.

[16] J. A. C. Guimarães. Abordagens teóricas em tratamento temático da informação: catalogação de assunto, indexação e análise documental. In: García Marco FJ (org.). *Avances y perspectivas en sistemas de información y de documentación.* Zaragoza: Prensas Universitarias de Zaragoza, 2009, pp. 105–117.

[17] B. Haslhofer, A. Isaac. *data.europeana.eu: The Europeana Linked Open Data Pilot.* 2011.

[18] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi. The curious case of neural text degeneration. In: *Proceedings of ICLR 2020.* 2019. URL: https://arxiv.org/abs/1904.09751.

[19] F. W. Lancaster. *Indexação e resumos: teoria e prática.* 3. ed. Brasília: Briquet de Lemos. 2004.

[20] A. L. M. Lemos. Erros, falhas e perturbações digitais em alucinações das IA generativas: tipologia, premissas e epistemologia da comunicação. *MATRIZes*, 18(1). 2024, pp. 75-91. doi: 10.11606/issn.1982-8160.v18i1p75-91.

[21] G. D. Maimone, L. S. Gracioso. Representação temática de imagens: perspectivas metodológicas. *Informação & Informação*, Londrina, v. 12, n. 1. 2007, pp. 11–28. URL: https://www.uel.br/revistas/uel/index.php/informacao/article/view/2764.

[22] C. H. Marcondes. Interoperabilidade entre acervos digitais de arquivos, bibliotecas e museus: potencialidades das tecnologias de dados abertos interligados. *Perspectivas em Ciência da Informação,* Belo Horizonte, v. 21, n. 2, 2016, pp. 61-83.

[23] M. P. Manini. *Análise documentária de fotografias: um referencial de leitura de imagens fotográficas para fins documentários.* 2002. 231 f. Tese (Doutorado em Ciências da Comunicação) – Universidade de São Paulo (USP), Escola de Comunicação e Artes. São Paulo, 2002.

[24] D. L. Martins, D. L. S. Lemos, L. F. R Oliveira, J. Siqueira, D. Carmo, V. N. Medeiros. Information organization and representation in digital cultural heritage in Brazil: systematic mapping of information infrastructure in digital collections for data science applications. *Journal of the Association for Information Science and Technology*, v. 74, n. 6, 2022, pp. 707–726. Doi: 10.1002/asi.24650.

[25] A. C. S. Oliveira. *Modelagem para valoração de objetos museológicos: estudos de caso para o MNBA e o MAST.* 2020. 1473 f. Tese (Doutorado em Museologia e Patrimônio) - Universidade

Federal do Estado do Rio de Janeiro (UNIRIO), Museu de Astronomia e Ciências Afins, Programa de Pós- Graduação em Museologia e Patrimônio. Rio de Janeiro, 2016.

[26] E. Panofsky. *Significado nas Artes Visuais.* 3. ed. São Paulo: Perspectiva, 1986.

[27] E. Panofsky. *Estudos de iconologia: temas humanísticos na arte do Renascimento.* Lisboa: Estamnpa, 1986.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Tech Rep.* 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[29] R. A. S. Ramalho. *Desenvolvimento e utilização de ontologias em bibliotecas digitais: uma proposta de aplicação.* 2010. 145 f. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista, Faculdade de Filosofia e Ciências. Marília, 2010.

[30] D. L. Silva, R. R. Souza. Representação de documentos multimídia: dos metadados às anotações semânticas. *Tendências da Pesquisa Brasileira em Ciência da Informação*, [S.l.], v. 7, n. 1, jan./jun. 2014.

[31] G. R. Silva. *Modelo de leitura para indexação de fotografias baseado no método complexo e nas funções primárias da imagem* [dissertação]. Belo Horizonte: Universidade Federal de Minas Gerais, Escola de Ciência da Informação; 2018. 140 f. URL: http://hdl.handle.net/1843/ECIP-B6VM9Y.

[32] P. N. Silva, C. C. Dias. Representação do conhecimento em tempos de inteligência artificial. In: ENCONTRO EDICIC, 14., 2024, Lisboa. *Diálogos na Ciência da Informação: atas do XIV Encontro EDICIC.* Lisboa: Centro de Estudos Clássicos, Faculdade de Letras, Universidade de Lisboa & Edições Colibri, jul. 2024.

[33] J. Smit. A representação da imagem. *Informare: Caderno do Programa de Pós-Graduação em Ciência da Informação*, Rio de Janeiro, 1996, pp. 28–36.

[34] F. Souza, R. Nogueira, R. Lotufo. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: *Intelligent Systems: 9th Brazilian Conference*, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I. Springer International Publishing, 2020, pp. 403-417. doi: 10.1007/978-3-030-61377-8_28.

[35] H. T. Sousa, G. A. Lima. Sumarização Automática de Textos Científicos com Inteligência Aumentada: Integrando IA na Organização do Conhecimento. *ISKO Brasil.* 2025. URL: https://isko.org.br/ojs/index.php/iskobrasil/article/view/44.

[36] A. A. L. Torres. *Metodologia para a representação de registro fotográfico de esculturas de arte sacra* [dissertação]. Belo Horizonte: Universidade Federal de Minas Gerais, Escola de Ciência da Informação; 2019. 205 f. Orientadora: B. C. M. S. Maculan. URL: https://repositorio.ufmg.br/handle/1843/31650.

[37] A. A. L. Torres, B. C. M. S. Maculan. Aspectos constitutivos e atributos da fotografia documental. In: N. B. Tognoli, A. C. Albuquerque, B. M. N. Cervantes (org). *Organização e representação do conhecimento em diferentes contextos: desafios e perspectivas na era da datificação.* Londrina: ISKO-Brasil, PPGCI-UEL. v. 1, 2023, p. 143–151.

[38] A. A. L. Torres, B. C. M. S. Maculan. A. A. Rocha, F. M. Assunção, F. B. Marques, G. S. Rodrigues. Inteligência artificial e a indexação de imagens com o método iconográfico de Panofsky. In: *Anais do VIII ISKO Brasil*; 2025; Canela, RS. v. 8. p. 1–14.

[39] M. L. Zeng. Semantic enrichment for enhancing LAM data and supporting digital humanities. *El profesional de la información*, v. 28, n. 1, e280103, 2019. Doi: 10.3145/epi.2019.ene.03.

[40] Y. Zhang, S. Sun, M. Galley, Y. C. Chen, C. Brockett, X. Gao, et al. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In: *Proc ACL 2020.* URL: https://arxiv.org/abs/1911.00536.

Sobre a Interação entre Inteligência Artificial Explicável (XAI) e Ontologias: Uma Revisão Quasi-Sistemática da Literatura

Lucas G. Maddalena[1,*], Fernanda Baião[1], Tiago Prince Sales[2] and Giancarlo Guizzardi[2]

[1]*Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rua Marquês de São Vicente, 225, Rio de Janeiro, 22451-900, Brasil*

[2]*Semantics, Cybersecurity & Services, University of Twente, Enschede, The Netherlands*

Abstract

Explainable Artificial Intelligence (XAI) seeks to reconcile the predictive power of modern Artificial Intelligence (AI) models—typically based purely on data-driven learning—with the human need for transparent and trustworthy decisions. This paper presents a quasi-systematic literature review, grounded in the PRISMA protocol principles [1], which examines the interplay between ontologies—formal, machine-readable representations of domain knowledge—and contemporary XAI techniques. A total of 440 records were retrieved from the Scopus database (2018–2025), which were screened and filtered according to predefined inclusion criteria. The final corpus was complemented by three seminal works recognized for the depth of their conceptual contributions, totaling 31 articles. The analysis aims to address three central questions: (1) how the notion of "explanation" is defined and operationalized in the XAI literature; (2) in which ways ontologies are employed as semantic artifacts in XAI systems—from term mapping to counterfactual reasoning; and (3) what impact ontological grounding has on the expressiveness of explanations and on user understanding. The results indicate that the concept of explanation is addressed in a heterogeneous manner—encompassing mechanistic, contrastive, and social perspectives—while ontologies are often used for domain-specific visualization and contextualization, with foundational ontologies still being underexplored. Ontological anchoring of explanations tends to enrich their semantic depth and shows potential to enhance user understanding, although empirical validation remains limited. The paper concludes with a discussion of emerging neuro-symbolic approaches and suggests future directions for integrating well-founded ontologies into explainable AI frameworks.

Keywords

Explainable Artificial Intelligence, Ontologies, XAI, Foundational Ontologies, Explainability, Literature Review

1. Introdução

Nos últimos anos, o campo da Inteligência Artificial (IA) tem vivenciado um notável aumento de popularidade. Embora existam várias definições sobre o que realmente constitui IA, adota-se neste trabalho a definição de Sheikh et al., que a descreve como sistemas que exibem comportamento inteligente ao analisar seu ambiente e tomar ações — com algum grau de autonomia — para alcançar objetivos específicos [2].

Contudo, as abordagens modernas de IA — especialmente aquelas baseadas em Aprendizado de Máquina (*Machine Learning*) e, em particular, em Aprendizado Profundo (*Deep Learning*) — são essencialmente orientadas por dados, baseando-se quase exclusivamente em padrões extraídos de grandes volumes de dados, que são considerados uma inteligência de conhecimento sub-simbólico, em detrimento da utilização de conhecimento externo codificado de forma explícita. Neste contexto, é notório que tais padrões apreendidos podem ser enviesados, não representativos e desprovidos de ancoragem no mundo real. Como resultado, a última década testemunhou tanto a ascensão das Redes Neurais

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Autor correspondente.

lucasmadda@aluno.puc-rio.br (L. G. Maddalena); fbaiao@puc-rio.br (F. Baião); t.princesales@utwente.nl (T. P. Sales); g.guizzardi@utwente.nl (G. Guizzardi)

0000-0001-6411-4259 (L. G. Maddalena); 0000-0001-7932-7134 (F. Baião); 0000-0002-5385-5761 (T. P. Sales); 0000-0002-3452-553X (G. Guizzardi)

Profundas — arquiteturas com centenas de camadas e milhões de parâmetros — quanto uma conscientização crescente de que tais arquiteturas resultam em modelos opacos, caracterizando-os como sistemas do tipo “caixa-preta” [3]. Diante desse cenário, a Inteligência Artificial Explicável (XAI) tem se consolidado como uma demanda central na academia e na indústria. À medida que sistemas de IA são aplicados em domínios críticos — como Saúde, Finanças, Sistemas Autônomos e Segurança —, torna-se essencial garantir que as partes interessadas compreendam o raciocínio subjacente dos modelos, construam confiança nos resultados produzidos e avaliem a robustez das previsões frente a entradas fora da distribuição, isto é, situações em que o modelo recebe dados substancialmente diferentes daqueles usados em seu treinamento. Exemplos típicos incluem exames clínicos realizados com equipamentos distintos ou em populações de pacientes com características demográficas ausentes no conjunto original de dados. Nesses casos, mesmo modelos altamente precisos podem apresentar degradação de desempenho ou produzir saídas enganosas, evidenciando a necessidade de explicações que revelem suas limitações. No entanto, embora essa necessidade venha sendo amplamente reconhecida, ainda há considerável divergência sobre o que constitui uma explicação válida. Em muitos casos, os métodos de XAI revelam apenas correlações estatísticas aprendidas a partir dos dados de treinamento, sem oferecer verdadeiro entendimento sobre o funcionamento do modelo, podendo inclusive reforçar vieses cognitivos ao fornecer justificativas superficiais.

O conhecimento externo explicitamente estruturado pode ser representado por ontologias — formalmente definidas como “uma especificação formal e explícita de uma conceitualização compartilhada” [4] — as quais oferecem uma fonte complementar de informação aos métodos puramente baseados em dados. Enquanto grande parte das técnicas de XAI se limita à visualização de padrões extraídos dos dados, as ontologias permitem fundamentar o comportamento dos modelos em conhecimento estruturado e legível por máquina, com potencial para mitigar vieses, ampliar a robustez preditiva e produzir explicações mais informativas e semanticamente válidas. Ao explicitar relações, categorias ontológicas e restrições do domínio, essas estruturas favorecem a geração de explicações alinhadas com o mundo real, promovendo maior transparência e apoio à decisão. Nesse sentido, a integração de modelos conceituais e aprendizado de máquina — como discutido em [5] — oferece uma via promissora para alinhar abstrações semânticas com ciclos de desenvolvimento em ciência de dados. Essa sinergia entre modelagem ontológica e aprendizado de máquina contribui não apenas para aumentar a expressividade das explicações, mas também para ancorá-las em representações cognitivamente adequadas e relevantes para os usuários finais.

Embora já existam revisões relevantes sobre o tema, como a de Ali et al. [6], que apresenta um mapeamento abrangente das técnicas e métricas de XAI, e a de Rajabi and Etminani [7], que sistematiza o uso de grafos de conhecimento em diferentes estágios de modelos explicáveis, ambas seguem direções distintas da proposta aqui desenvolvida. A primeira centra-se em categorizar métodos e ferramentas disponíveis para XAI em geral, sem adentrar especificamente no papel de ontologias; a segunda foca principalmente em grafos de conhecimento, mencionando ontologias apenas de forma tangencial, sem discutir sua distinção conceitual nem seu potencial quando fundamentadas em ontologias de topo. Em contraste, o presente trabalho busca examinar como ontologias — enquanto artefatos semânticos formais e bem fundamentados — têm sido integradas a sistemas de XAI, enfatizando seus efeitos na qualidade e expressividade das explicações. Dessa forma, nossa revisão avança em uma direção ainda pouco explorada, complementando os panoramas já existentes e preenchendo uma lacuna sobre o uso de ontologias propriamente ditas na explicabilidade de modelos de IA.

Com base nessa perspectiva, este trabalho realiza uma revisão quasi-sistemática da literatura com o objetivo de examinar como ontologias e outras formas de representação do conhecimento vêm sendo incorporadas a técnicas contemporâneas de XAI, com ênfase especial em abordagens neurossimbólicas que integram raciocínio simbólico e aprendizado sub-simbólico, e endereçando três questões centrais de pesquisa: (QP1) Como o termo “explicação” é definido e operacionalizado na literatura atual sobre XAI?; (QP2) De que formas as ontologias são utilizadas como artefatos semânticos em sistemas de XAI?; e (QP3) Como a fundamentação de explicações em ontologias afeta sua expressividade e a compreensão por parte dos usuários?

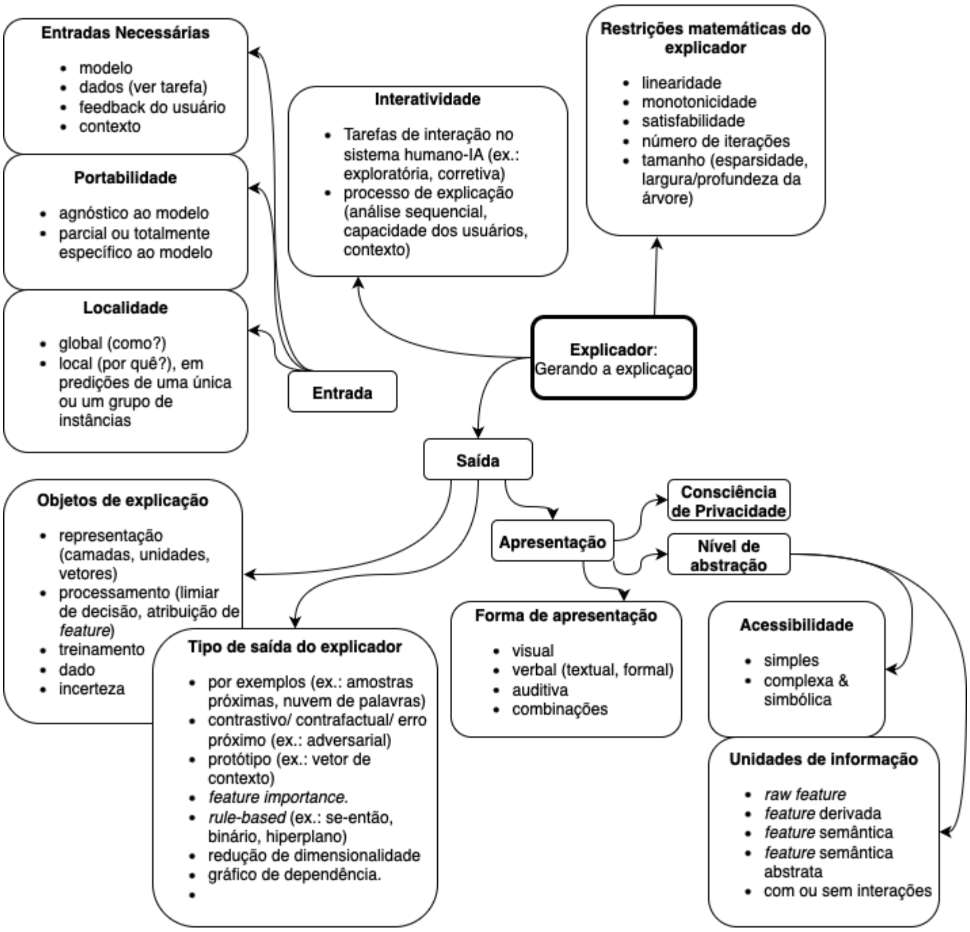
Inicialmente, o artigo delimita os conceitos fundamentais que sustentam a pesquisa, com destaque para a Inteligência Artificial Explicável, a função das ontologias como artefatos semânticos e os paradigmas emergentes de IA neurossimbólica. Essa fundamentação teórica estabelece o alicerce para a metodologia apresentada na Seção 3, na qual são descritos o desenho da pesquisa, os critérios de inclusão e exclusão, a estratégia de busca utilizada e os procedimentos de triagem e seleção bibliográfica. A Seção 4 sintetiza os principais achados da revisão, identificando padrões recorrentes e lacunas existentes na literatura sobre a integração entre ontologias e XAI. São discutidos criticamente os tipos de ontologias utilizados, os mecanismos de explicação empregados, bem como os efeitos observados na transparência e na compreensão por parte dos usuários. Por fim, a Seção 5 apresenta considerações finais sobre as implicações dos resultados obtidos e propõe direções promissoras para pesquisas futuras.

2. Referencial Teórico

A Inteligência Artificial Explicável (XAI) surgiu como resposta direta à crescente opacidade dos modelos modernos de aprendizado de máquina — particularmente das redes neurais profundas, cujas arquiteturas podem abranger centenas de camadas e milhões de parâmetros — tornando-os sistemas "caixa-preta" de difícil interpretação [3]. Paradigmas iniciais de IA, como os sistemas especialistas baseados em regras, forneciam caminhos decisórios inerentemente transparentes, mas as abordagens orientadas por dados abrem mão dessa clareza em prol do poder preditivo. Como observam Confalonieri et al. [8], essa mudança deu origem a preocupações cognitivas e sociais urgentes: os envolvidos precisam não apenas saber o que o modelo prediz, mas também entender o porquê — de modo que as explicações estejam alinhadas com noções humanas de causalidade e responsabilidade. Mais recentemente, Ali et al. [6] destacam que a XAI é fundamental para a construção de uma IA confiável (Trustworthy AI) — garantindo transparência, confiabilidade e justiça em domínios críticos como saúde e finanças. No entanto, como é alertado por Mittelstadt et al. [9], muitas das chamadas "explicações" são pouco mais que modelos substitutos simplificados, reaproveitados da modelagem científica, mas sem estarem fundamentados nas práticas contrastivas, seletivas e situadas socialmente que caracterizam as explicações genuinamente humanas.

Para trazer ordem a esse campo complexo, taxonomias de métodos XAI geralmente distinguem entre modelos ditos "intrinsecamente transparentes" e aqueles que exigem interpretação *post-hoc*. A primeira categoria — que inclui regressões (lineares, logísticas, entre outras), árvores de decisão e até mesmo ferramentas baseadas em modelos substitutos como LIME [10] e SHAP [11] — oferece rastreabilidade direta por meio de coeficientes ou regras explícitas, mas ainda assim permanece totalmente orientada por dados e vulnerável a extrapolações enganosas em entradas fora da distribuição [3]. Por outro lado, as técnicas *post-hoc* se dividem em abordagens independentes de modelo (modelo como caixa-preta) e métodos específicos de modelo que aproveitam detalhes arquiteturais (ex: mapas de saliência, propagação de relevância por camadas). As explicações também podem ser classificadas como locais — que elucidam previsões individuais — ou globais — que resumem o comportamento geral do modelo. Embora essas distinções auxiliem na escolha e avaliação dos métodos, Mittelstadt et al. [9] nos lembram que um foco estreito em aproximações funcionais corre o risco de negligenciar as dimensões normativas mais profundas da explicação — como a necessidade de contrafactuais contrastivos, relevância contextual e interação dialógica. Abordar essas lacunas é central para avançar a XAI de um conjunto de ferramentas exclusivamente técnicas para uma disciplina centrada no ser humano, que de fato capacite tanto usuários quanto reguladores.

Com base nessas fundações, Schwalbe and Finzel [12] propõem uma taxonomia abrangente de métodos XAI, ilustrada na 1, que organiza sistematicamente as características explicativas em dimensões inter-relacionadas. Partindo dos requisitos de entrada (incluindo acesso ao modelo, dados relevantes, feedback do usuário e informações contextuais), a taxonomia distingue os métodos por portabilidade (agnóstico vs. específico de modelo) e localidade (explicações globais versus locais). Em seguida, define critérios de interatividade (fluxos de trabalho exploratórios ou corretivos), restrições matemáticas (linearidade, monotonicidade, satisfatibilidade, limites de iteração e tamanho do modelo) e tipos de saída (baseados em exemplos, contrastivos/contrafactuais, guiados por protótipos, atribuições de características, baseados em regras, redução de dimensionalidade, gráficos/diagramas). Por fim, aborda fatores de apresentação — modalidade (visual, verbal, auditiva), nível de abstração, acessibilidade, granularidade da informação e consciência de privacidade — oferecendo, assim, um arcabouço estruturado para seleção e avaliação de técnicas XAI conforme os requisitos específicos de uso.



Entradas Necessárias
modelo
dados (ver tarefa)
feedback do usuário
contexto
Portabilidade
agnóstico ao modelo
parcial ou totalmente específico ao modelo
Localidade
global (como?)
local (por quê?), em predições de uma única ou um grupo de instâncias
Interatividade
Tarefas de interação no sistema humano-IA (ex.: exploratória, corretiva)
processo de explicação (análise sequencial, capacidade dos usuários, contexto)
Restrições matemáticas do explicador
linearidade
monotonicidade
satisfabilidade
número de iterações
tamanho (esparsidade, largura/profundeza da árvore)
Entrada
Explicador: Gerando a explicaçao
Saída
Objetos de explicação
representação (camadas, unidades, vetores)
processamento (limiar de decisão, atribuição de feature)
treinamento
dado
incerteza
Tipo de saída do explicador
por exemplos (ex.: amostras próximas, nuvem de palavras)
contrastivo/ contrafactual/ erro próximo (ex.: adversarial)
protótipo (ex.: vetor de contexto)
feature importance.
rule-based (ex.: se-então, binário, hiperplano)
redução de dimensionalidade
gráfico de dependência.
Apresentação
Consciência de Privacidade
Nível de abstração
Forma de apresentação
visual
verbal (textual, formal)
auditiva
combinações
Acessibilidade
simples
complexa & simbólica
Unidades de informação
raw feature
feature derivada
feature semântica
feature semântica abstrata
com ou sem interações

Figura 1: Aspectos taxonômicos de um explanador, conforme proposto por [12]

A partir dessa taxonomia e do reconhecimento das limitações das explicações puramente orientadas por dados, a IA Neurossimbólica surge como uma promissora terceira onda da Inteligência Artificial, unindo a capacidade de reconhecimento de padrões do aprendizado de máquina (profundo) ao rigor e interpretabilidade do raciocínio simbólico [13, 14]. Enquanto redes sub-simbólicas se destacam na extração de características latentes a partir de grandes conjuntos de dados, mas oferecem pouco em termos de justificativas compreensíveis para humanos, os sistemas neurossimbólicos incorporam representações formais de conhecimento — ontologias, regras lógicas ou restrições probabilísticas — diretamente nas arquiteturas neurais ou em paralelo a elas. Essa fusão fundamentada permite que os modelos realizem aprendizado por gradiente em tarefas perceptivas ao mesmo tempo em que sustentam inferência simbólica estruturada, raciocínio contrafactual contrastivo e generalização composicional. Ao unir esses paradigmas, a IA Neurossimbólica pretende combinar alto desempenho preditivo com caminhos decisórios intrinsecamente explicáveis, lançando as bases para abordagens XAI que realmente abordem as dimensões normativas, seletivas e interativas da explicação humana.

Essa convergência entre o rigor simbólico e o aprendizado sub-simbólico ressalta a necessidade de ontologias de fundamentação de alta qualidade como alicerce dos sistemas explicáveis. Estruturas como a Unified Foundational Ontology (UFO)[15], a DOLCE [16] e a Basic Formal Ontology (BFO) [17] oferecem conjuntos gerais de categorias e relações, apoiados em compromissos filosóficos claros e metodologias consolidadas. Ao ancorar o conhecimento de domínio nessas ontologias e ao empregar linguagens de modelagem conceitual como a OntoUML, garante-se que o componente simbólico seja simultaneamente semanticamente rico e formalmente consistente. Essas ontologias impõem restrições à inferência neural e possibilitam explicações rastreáveis a conceitos compartilhados e legíveis por humanos, reduzindo a lacuna de interpretabilidade inerente às abordagens exclusivamente orientadas por dados.

3. Metodologia

Este estudo adota uma abordagem de *revisão quasi-sistemática* da literatura, guiada por princípios do protocolo PRISMA [1], com o objetivo de assegurar transparência e reprodutibilidade na seleção dos trabalhos analisados. Foram empregadas boas práticas metodológicas amplamente reconhecidas, tais como: definição antecipada de critérios de inclusão e exclusão; construção de uma estratégia de busca estruturada; e registro visual do processo de triagem por meio de fluxograma. A revisão foi orientada por três perguntas centrais de pesquisa:

- **QP1**: Como o termo "explicação" é definido e operacionalizado na literatura atual sobre XAI?
- **QP2**: De que formas as ontologias são utilizadas como artefatos semânticos em sistemas de XAI?
- **QP3**: Como a fundamentação de explicações em ontologias afeta sua expressividade e a compreensão por parte dos usuários?

Com relação à QP1, buscou-se compreender, especificamente, quais dimensões de explicação são enfatizadas pelos autores e quão consistentes essas definições são entre domínios e venues. Essas dimensões incluem perspectivas mecanicistas (como entradas específicas afetam as saídas de um modelo), contrastivas (explicações do tipo "por que A e não B?") e sociais (que consideram o perfil, contexto e necessidades do usuário final) [12]. Para a QP2, investigou-se quais tipos de ontologias (de fundamentação vs. específicas de domínio) aparecem com maior frequência e quais papéis desempenham (ex.: fundamentação de termos, estruturação causal, raciocínio contrafactual). Finalmente, na QP3, buscou-se compreender se explicações baseadas em ontologias demonstram maior riqueza semântica (relações explícitas, restrições) e se há evidências empíricas de que os usuários as consideram mais transparentes ou acionáveis.

A base de dados Scopus foi consultada utilizando a seguinte string de busca: `TITLE-ABS-KEY(("explainable ai" OR xai OR explainability OR interpretability OR "neuro-symbolic") AND (ontology OR ontologies OR ontological OR "foundational ontology" OR ufo OR ontouml)).`

Com o intuito de concentrar a análise em contribuições mais consolidadas no campo da XAI, foram considerados apenas artigos publicados em sua versão final a partir de 2018 — marco temporal a partir do qual o termo "XAI" passou a figurar com maior destaque em títulos, resumos e palavras-chave.

A consulta inicial retornou 440 registros. Primeiramente, removemos 386 itens que não eram artigos completos (como cartas, atas de conferências, editoriais), restando 54 artigos para triagem por título, resumo e palavras-chave. Nessa fase, excluímos 17 estudos sem ligação substancial com IA explicável, ontologias ou sistemas de representação do conhecimento, além de outros 9 artigos que mencionavam XAI apenas de forma tangencial, sem discussão explícita sobre mecanismos de explicação, fundamentos ontológicos ou integração sistemática de artefatos semânticos. Isso resultou em 28 artigos selecionados após a triagem.

Como mostrado na Figura 2, recuperamos 440 registros da Scopus, dos quais 386 foram removidos por não serem pertinentes. Os 54 restantes foram avaliados com base em título, resumo e, quando necessário, no texto completo; nessa etapa, 17 foram excluídos por ausência de vínculo com XAI ou ontologias, e 9 por não abordarem mecanismos de explicação ou fundamentos ontológicos. Reconhecendo, entretanto, que nossas questões de pesquisa demandavam uma análise aprofundada das estruturas ontológicas aplicadas em XAI, o corpus foi complementado com três trabalhos seminais previamente conhecidos por suas definições rigorosas e tratamento detalhado do papel das ontologias na explicação [18, 19, 20].

Assim, a amostra final totalizou 31 artigos, submetidos à extração e síntese detalhadas para responder às três perguntas de pesquisa propostas.

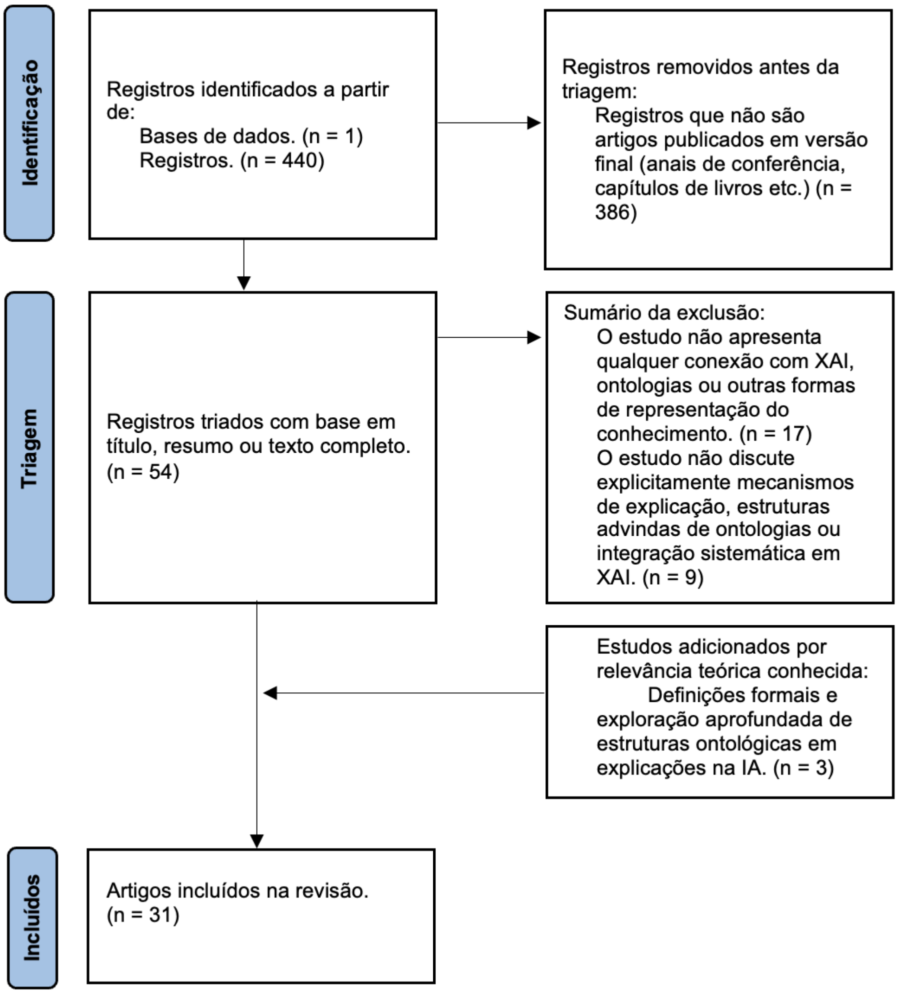

Identificação
Registros identificados a partir de:
Bases de dados. (n = 1)
Registros. (n = 440)
Registros removidos antes da triagem:
Registros que não são artigos publicados em versão final (anais de conferência, capítulos de livros etc.) (n = 386)
Triagem
Registros triados com base em título, resumo ou texto completo. (n = 54)
Sumário da exclusão:
O estudo não apresenta qualquer conexão com XAI, ontologias ou outras formas de representação do conhecimento. (n = 17)
O estudo não discute explicitamente mecanismos de explicação, estruturas advindas de ontologias ou integração sistemática em XAI. (n = 9)
Estudos adicionados por relevância teórica conhecida:
Definições formais e exploração aprofundada de estruturas ontológicas em explicações na IA. (n = 3)
Incluídos
Artigos incluídos na revisão. (n = 31)

Figura 2: Fluxograma do processo de seleção dos estudos. Fonte: Elaborado pelos autores.

4. Resultados

A Tabela 1 resume as publicações selecionadas e suas respectivas contribuições para cada uma das três perguntas de pesquisa (QP1: definição e operacionalização de "explicação" no contexto de XAI; QP2: papel das ontologias como artefatos semânticos; e QP3: impacto da fundamentação ontológica na expressividade das explicações e na compreensão por parte dos usuários), indicando se a publicação apresentou contribuição relevante ("Sim") ou não ("Não") e assinalando com "*" os casos em que observações ou nuances adicionais são discutidas posteriormente. Conforme evidenciado, alguns trabalhos, como [18, 19, 20], oferecem contribuições significativas para todas as perguntas de pesquisa, ao passo que estudos como [21] e [22] não abordam diretamente nenhuma das questões delineadas.

4.1. Análise da QP1: Definição e operacionalização de "explicação"em XAI

A maioria das publicações selecionadas aborda, de forma direta ou indireta, a primeira questão de pesquisa (QP1), relacionada à definição e operacionalização do conceito de explicação em Inteligência Artificial Explicável (XAI). Entretanto, alguns estudos apresentam limitações quanto à profundidade dessa discussão. Por exemplo, Sun et al. [21] apenas menciona a falta de interpretabilidade em modelos

Tabela 1
Artigos selecionados e contribuições para as Questões de Pesquisa endereçadas no presente estudo.

Citação	Referência	Ano	QP1	QP2	QP3
[18]	Guizzardi and Guarino	2024	Sim	Sim	Sim
[19]	Confalonieri and Guizzardi	2023	Sim	Sim	Sim
[20]	Chari et al.	2020	Sim	Sim	Sim
[21]	Sun et al.	2022	Não	Não	Não
[22]	Lu et al.	2023	Sim	Não	Não
[23]	Li et al.	2025	Sim	Sim*	Não
[24]	Liu et al.	2022	Sim	Sim	Sim
[25]	Perera and Liu	2024	Sim	Sim*	Não*
[26]	Ma et al.	2025	Sim*	Sim*	Não
[27]	Wei et al.	2024	Sim*	Não	Não
[28]	Zhao et al.	2024	Sim	Sim	Sim
[29]	Rasbach et al.	2024	Sim	Não*	Não
[30]	Dubey et al.	2025	Sim	Não	Não
[31]	Zhao et al.	2022	Sim	Sim	Não
[32]	Zhan et al.	2025	Sim	Sim	Sim
[33]	Sung et al.	2024	Sim	Sim	Sim
[34]	Li et al.	2025	Sim	Sim	Sim
[35]	Zheng et al.	2024	Sim	Sim	Não
[36]	Chen et al.	2018	Não*	Sim*	Sim*
[37]	Canito et al.	2021	Sim*	Sim*	Sim*
[38]	Meghraoui et al.	2025	Não*	Sim	Sim
[39]	Guan et al.	2024	Sim	Não	Não
[40]	Basakin et al.	2024	Sim	Sim	Sim
[41]	Li et al.	2023	Sim	Não	Sim
[42]	Kaptein et al.	2022	Sim	Sim	Não
[43]	Smedley et al.	2021	Sim	Sim	Sim
[44]	Ruiz-Arenas et al.	2024	Sim	Sim	Não
[45]	Seninge et al.	2021	Sim	Sim	Sim
[46]	Singha et al.	2023	Sim	Não	Não
[47]	Askland et al.	2021	Sim*	Sim	Sim
[48]	Townsend et al.	2024	Sim	Não	Sim
		Total de "Sim"	**27**	**23**	**17**

de IA do tipo caixa-preta, sem aprofundar nos tipos de explicação ou modelos adotados na literatura. De forma semelhante, Ma et al. [26] reconhece a importância da interpretabilidade, mas não a relaciona com as dimensões conceituais da explicação tratadas no campo de XAI. O estudo de Wei et al. [27], embora utilize recursos funcionais para predizer variantes genéticas patogênicas, também não define explicitamente o que se entende por "explicação" nem discute suas possíveis dimensões.

Já os trabalhos de Chen et al. [36] e Meghraoui et al. [38] exemplificam abordagens que, embora não se identifiquem diretamente com o campo de XAI como hoje é concebido, contribuem para a interpretabilidade dos modelos por meio da incorporação de conhecimento estruturado. No primeiro caso, são integrados recursos como WordNet [49] e UMLS (Unified Medical Language System) [50] às análises de embeddings. No segundo, adota-se uma abordagem neurossimbólica aplicada à previsão de produtividade agrícola, cuja natureza simbólica intrínseca tende a favorecer a explicabilidade do sistema, ainda que sem uma formalização explícita do conceito de explicação.

No estudo de Canito et al. [37], é apresentada uma revisão sistemática sobre evolução ontológica com restrição temporal no domínio de manutenção preditiva. Embora não esteja diretamente inserido no escopo da XAI, o trabalho oferece uma contribuição relevante à operacionalização de explicações em contextos que envolvem conhecimento estruturado e dinâmico ao longo do tempo. O modelo proposto para representar a evolução de ontologias com validade temporal relaciona-se ao desafio de justificar e rastrear mudanças no comportamento de sistemas inteligentes, o que é essencial para explicações retrospectivas e prospectivas. A marcação (*) na Tabela 1, nesse caso, reflete o fato de tratar-se de uma revisão sistemática, cuja abrangência metodológica contribui indiretamente para a semântica da explicação por meio de mecanismos ontológicos, no caso em questão sensíveis especificamente ao aspecto temporal.

Já no trabalho de Askland et al. [47], embora não haja definição ou operacionalização explícita de "explicação" nas dimensões mais comuns da literatura — como as perspectivas mecanicista, contrastiva ou social —, identifica-se uma abordagem implícita de cunho mecanicista, dada a ênfase na contextualização biológica e nas interações multinível na análise de dados genéticos.

Em síntese, observa-se uma diversidade de abordagens em relação à QP1. Parte dos estudos ([18, 19, 20]) se destaca por apresentar uma fundamentação teórica robusta e abrangente, contemplando dimensões formais da explicação, como o raciocínio contrastivo e mecanicista. Outros trabalhos ([37, 47]) oferecem contribuições mais periféricas, mas ainda relevantes, ao introduzirem elementos estruturantes — temporais ou biológicos — que favorecem a rastreabilidade e contextualização das inferências. Em contrapartida, estudos como [21, 26, 27] abordam a interpretabilidade de forma genérica, sem estabelecer vínculos explícitos com o arcabouço teórico da XAI. De modo geral, a literatura revisada oferece uma resposta heterogênea, porém majoritariamente positiva, à primeira pergunta de pesquisa, revelando diferentes graus de alinhamento com os referenciais conceituais mais consolidados da área.

4.2. Análise da QP2: De que formas as ontologias são utilizadas como artefatos semânticos em sistemas de XAI?

A análise das publicações selecionadas revela um amplo espectro de formas pelas quais ontologias são empregadas em sistemas de XAI, variando significativamente quanto ao nível de expressividade semântica. Os estudos de Liu et al. [24] e Kaptein et al. [42] apresentam evidências claras do uso de ontologias tanto de topo quanto específicas de domínio, utilizando-as para fundamentação de termos, estruturação causal e, em certos casos, permitindo uma integração modular que pode sustentar raciocínios contrafactuais. Essas contribuições alinham-se diretamente aos papéis previstos na QP2 e foram classificadas como evidências fortes na Tabela 1.

Outros trabalhos, como [26, 28, 29], recorrem majoritariamente a ontologias de domínio — como MedDRA [51] e Gene Ontology (GO) [52] — com o objetivo de enriquecer representações semânticas e aprimorar a interpretabilidade biológica ou clínica. Embora essas implementações fortaleçam a ancoragem semântica dos termos utilizados, raramente articulam uma estrutura mais abrangente para raciocínios causais ou contrafactuais, além de frequentemente deixarem de fazer uso de ontologias bem fundamentadas em uma ontologia de topo.

No caso de Perera and Liu [25], o uso de ontologias é apenas implícito, inserido em abordagens mais amplas baseadas em LLMs para aprendizado ontológico — uma área adjacente relevante, mas que não trata da utilização operacional de ontologias em pipelines de XAI. De forma semelhante, Li et al. [53] e Chen et al. [36] empregam embeddings e artefatos semânticos derivados de recursos estruturados (como WordNet [49], UMLS [50] e ProteinBERT [54]), contribuindo indiretamente para a fundamentação semântica, mas sem o emprego explícito de engenharia ontológica formal. Ainda que esses trabalhos não se enquadrem estritamente no escopo da XAI, a incorporação de conhecimento estruturado ao longo do pipeline de aprendizado de modelos de IA — mesmo que de forma parcial — representa um avanço rumo a sistemas semanticamente comprometidos, potencialmente mais alinhados com o mundo real e com maior capacidade de interpretação contextualizada, conforme apontaram Maass and Storey [55] e Amaral et al. [56]. Destacam-se também os estudos de Meghraoui et al. [38] e Basakin et al. [40], que apresentam abordagens neurossimbólicas integradas, demonstrando como a combinação de modelos baseados em conhecimento e aprendizado estatístico pode ser operacionalizada por meio do uso explícito de ontologias e técnicas modernas de criação de *embeddings*. Em particular, Meghraoui et al. [38] propõem uma ontologia dedicada à predição de produtividade agrícola, cujo conteúdo é formalmente avaliado e integrado a modelos de *machine learning*, ilustrando como restrições ontológicas podem sustentar raciocínios mais robustos e interpretáveis. Já os trabalhos de Zhan et al. [32] e Sung et al. [33] exploram a ancoragem ontológica como forma de aprimorar a transparência e a contextualização das explicações, mesmo sem recorrer a ontologias de fundamentação. Em [32], a ontologia é utilizada para estruturar mecanismos explicativos em medicina de precisão, conectando entidades biomédicas a raciocínios inferenciais em sistemas de recomendação terapêutica. De forma complementar, em [33] os autores aplicam uma ontologia modular para segmentar e representar raciocínios diagnósticos, promovendo explicações mais compreensíveis em ambientes clínicos. Em ambos os casos, destaca-se o compromisso semântico estabelecido pelas ontologias no pipeline de IA, o que aproxima os modelos resultantes do raciocínio humano e do conhecimento do domínio. Por sua vez, Basakin et al. [40] enfatizam o papel de ontologias em tarefas de tomada de decisão médica, utilizando-as para representar conhecimentos biomédicos que orientam a geração de explicações simbólicas alinhadas com dados clínicos.

Por outro lado, estudos como [47, 21] empregam ontologias específicas de domínio principalmente como ferramentas para seleção de atributos ou contextualização biológica, sem recorrer à estruturação de inferências explicáveis.

A contribuição seminal de Confalonieri and Guizzardi [19] oferece uma perspectiva profunda e teoricamente fundamentada sobre a integração de ontologias em sistemas de XAI, ao apresentar uma taxonomia abrangente dos papéis que esses artefatos podem desempenhar na construção de explicações. Essa taxonomia está organizada em três dimensões principais: a primeira, voltada à modelagem de referência, destaca o uso de ontologias como base conceitual para representar de forma precisa e compartilhável o domínio de interesse, servindo como estrutura semântica para a geração de explicações alinhadas ao conhecimento especializado; a segunda, centrada no raciocínio de senso comum, explora o papel das ontologias na sustentação de inferências que extrapolam os dados disponíveis, promovendo explicações mais compreensíveis e contextualizadas para usuários não técnicos; e a terceira, relativa ao refinamento do conhecimento, trata do uso de ontologias para revisar, modular e adaptar o conteúdo explicativo, viabilizando a personalização das justificativas e o aprimoramento contínuo dos modelos explicáveis. O estudo articula com clareza como essas dimensões tornam as ontologias componentes essenciais na gestão da complexidade, na adaptação das explicações a diferentes perfis de usuários e na viabilização de raciocínios automatizados mais transparentes e robustos.

Por fim, o estudo feito em [18] leva o conceito de explicação a um novo patamar ao propor um arcabouço sistemático para explicações ontologicamente fundamentadas em sistemas de IA. O trabalho está ancorado na abordagem de *Ontology-Driven Conceptual Modeling* (ODCM), que busca explorar o papel das ontologias no suporte à construção, análise e interpretação de modelos conceituais. Embora a ODCM possa se apoiar em diferentes tipos de ontologias, o estudo em questão emprega a OntoUML — uma reformulação da linguagem UML cujos construtores sintáticos e semânticos derivam diretamente das distinções ontológicas e axiomatizações propostas pela ontologia de topo UFO (Unified Foundational Ontology) [15]. Essa integração possibilita a exploração plena de artefatos semânticos ricos, como os estereótipos ontológicos de OntoUML (e.g., *Kind*, *Role*, *Relator*), as cardinalidades fundamentadas em relações ontológicas e os chamados *ontological design patterns*, que capturam regularidades estruturais recorrentes. Tais elementos não apenas conferem maior expressividade e rigor formal aos modelos, como também servem de base para uma abordagem explicativa mais profunda.

Nesse contexto, destaca-se a técnica de *ontological unpacking*, que busca explicar descrições simbólicas — como modelos conceituais, grafos de conhecimento ou especificações formais — revelando seus compromissos ontológicos subjacentes. Isso é feito por meio da identificação de *truthmakers*, ou seja, as entidades do domínio que tornam verdadeiras as proposições modeladas (*truthbearers*). Ao tornar explícita essa camada ontológica, o processo explicativo se torna mais robusto, estruturado e cognitivamente adequado, indo além de justificativas superficiais para incorporar elementos de causalidade, dependência e temporalidade de forma logicamente coerente. Essa proposta representa um avanço significativo no uso de ontologias como suporte à geração de explicações verdadeiramente semânticas e alinhadas com os fenômenos do mundo real.

4.3. Análise da QP3: Como a fundamentação de explicações em uma ontologia afeta sua expressividade e a compreensão por parte dos usuários?

Fundamentar explicações em uma ontologia de fundamentação de forma consistente amplifica sua riqueza semântica ao tornar explícitas as relações e restrições envolvidas e melhora sua clareza e capacidade de ação por parte dos usuários finais. Liu et al. [24] demonstram que o paradigma de modelagem multinível gera explicações mais expressivas, uma vez que a estrutura em "cubo conceitual" revela dependências latentes e impõe coerência semântica, indicando maior interpretabilidade, ainda que estudos com usuários permaneçam pendentes. De forma semelhante, Zhao et al. [28] incorpora anotações da Gene Ontology ao seu visualizador explicativo *Gradient Weighted interaction Activation Mapping* (Grad-WAM), destacando sítios funcionalmente relevantes de aminoácidos e vinculando atribuições a termos estabelecidos do domínio; embora avaliações formais com usuários estejam ausentes, os vínculos com a GO oferecem um caminho claro rumo à transparência. Em Zhan et al. [32], relata-se que as explicações fundamentadas na ontologia *Myocardial infarction Ontology* (Mio) permitem inferência automatizada de conhecimento médico, tornando-as simultaneamente mais transparentes e acionáveis, enquanto Sung et al. [33] mostra que o uso da *Kyoto Encyclopedia of Genes and Genomes* (KEGG) como base ontológica em seu framework *Multi-Dimensional Transcriptomic Ruler* (MDTR) gera insights estruturados e multidimensionais sobre mecanismos de hepatotoxicidade — características que, em princípio, favorecem a compreensão do usuário ao conectar as saídas do modelo a vias biológicas bem compreendidas.

Nos casos em que avaliações empíricas foram conduzidas, os resultados confirmam o valor das explicações baseadas em ontologias. O trabalho de Smedley et al. [43] demonstra, por meio de comparações de AUC, que mascarar atributos segundo conjuntos gênicos Hallmark e GO não apenas preserva a fidelidade preditiva, como também gera explicações alinhadas ao conhecimento biológico já estabelecido, indicando maior confiança por parte dos usuários. Em Seninge et al. [45], avança-se ainda mais, mostrando que a arquitetura *Variational Autoencoder Enhanced by Gene Annotations* (VEGA) — cuja estrutura de decodificação reflete módulos gênicos definidos por especialistas — permite a construção de espaços latentes interpretáveis e gera insights acionáveis sobre a atividade de vias; feedback preliminar de usuários sugere que essas visualizações orientadas por ontologias de fato promovem maior transparência. Em contraste, Perera and Liu [25] observam que os métodos atuais baseados em LLMs (como mapas de atenção ou encadeamento de raciocínios — *Chain-of-Thought*) permanecem "longe de resolvidos" para aprendizado ontológico, ressaltando a necessidade de arcabouços semânticos mais robustos. Os trabalhos de Canito et al. [37] e Meghraoui et al. [38] reconhecem que, apesar de seus sistemas incorporarem conhecimento rico de domínio e temporalidade, ainda faltam estudos sistemáticos com usuários que avaliem a compreensão — apontando para uma lacuna relevante.

Por fim, revisões de escopo mais amplo e análises conceituais corroboram esses achados. Por exemplo, Basakin et al. [40] argumentam que explicações fundamentadas em ontologias — ao associar a semântica dos dados das saídas do modelo — tendem a promover maior confiança por parte dos usuários, mesmo que sua validação empírica ainda esteja em aberto. De forma mais aprofundada, Guizzardi and Guarino [18] propõem o uso de *ontological unpacking* como um processo explicativo que visa tornar explícitos os compromissos ontológicos embutidos em artefatos simbólicos (como modelos conceituais, grafos de conhecimento ou especificações lógicas). Essa abordagem permite revelar os *truthmakers* — entidades ontológicas que fundamentam a veracidade de uma afirmação — e sustentar explicações contrastivas, contrafactuais e logicamente coerentes, por meio da exploração sistemática dos elementos fornecidos por uma ontologia bem fundamentada, como os estereótipos ontológicos da UFO, padrões ontológicos e restrições de cardinalidade.

5. Conclusões

Embora esta pesquisa tenha seguido princípios do protocolo PRISMA 2020, trata-se de uma revisão quasi-sistemática. Isso significa que, embora tenha adotado um processo estruturado de busca, critérios de elegibilidade previamente definidos e uma análise metodológica dos achados, a revisão não possui o caráter de exaustividade de uma revisão sistemática completa. Uma limitação relevante é que a busca foi conduzida exclusivamente na base Scopus, o que pode ter restringido a abrangência do corpus e deixado de fora trabalhos relevantes indexados em outras bases. Ainda assim, o procedimento empregado garantiu transparência, rastreabilidade e alinhamento com as boas práticas recomendadas pelo PRISMA, assegurando a consistência necessária para responder às questões de pesquisa propostas.

À luz dessas considerações, esta revisão quasi-sistemática analisou como ontologias têm sido empregadas para apoiar a Inteligência Artificial Explicável (XAI), especialmente sob perspectivas semânticas e epistêmicas. Os principais achados podem ser sumarizados a partir das perguntas de pesquisa originalmente formuladas:

QP1: Como o termo "explicação" é definido e operacionalizado na literatura atual sobre XAI?

As definições de explicação ainda são heterogêneas, com enfoques distintos em dimensões *mecanicistas* (como entradas afetam saídas), *contrastivas* (explicações "por que A e não B?"), e *sociais* (considerando o perfil e contexto do usuário). Contudo, poucos trabalhos apresentam abordagens sistematicamente fundamentadas. Apenas uma minoria, como Guizzardi and Guarino [18] e Confalonieri and Guizzardi [19], propõe modelos explicativos com base em teorias filosóficas ou ontológicas robustas.

QP2: De que formas as ontologias são utilizadas como artefatos semânticos em sistemas de XAI?

As ontologias são majoritariamente utilizadas para a *fundamentação de termos* e a *contextualização semântica*, especialmente em domínios biomédicos. Ontologias específicas de domínio, como Gene Ontology e KEGG, aparecem com mais frequência, mas seu uso tende a ser superficial. Ontologias de fundamentação, como a UFO, são raramente exploradas, apesar de seu potencial para apoiar *raciocínio causal*, *explicações contrafactuais* e garantir menor ambiguidade semântica. Trabalhos como de Guizzardi and Guarino [18] demonstram que o uso de modelos conceituais ontologicamente bem fundados permite maior expressividade e clareza nos comprometimentos semânticos das explicações.

QP3: Como a fundamentação de explicações em ontologias afeta sua expressividade e a compreensão por parte dos usuários?

A fundamentação ontológica tende a ampliar significativamente a *expressividade semântica* das explicações, tornando explícitas as relações e restrições entre conceitos. Visualizações baseadas em ontologias, como as que utilizam a Gene Ontology ou o KEGG, permitem associações mais ricas e estruturadas. Há indícios de que essas abordagens aumentam a *transparência percebida* e a *confiança* dos usuários. No entanto, a maior parte dos estudos carece de avaliações empíricas sistemáticas sobre a compreensão e a utilidade das explicações.

Com base nesses achados, delineiam-se algumas recomendações para pesquisas futuras. É necessário promover a **criação de ontologias de domínio fundamentadas em ontologias de fundamentação**, explorando distinções ontológicas bem estabelecidas para estruturar explicações contrastivas, causais e temporais. Ainda, é importante ampliar **estudos sistemáticos com usuários** que avaliem a compreensão, confiança e eficácia das explicações baseadas em ontologias; outras questões ainda em aberto incluem desenvolver **métricas e benchmarks padronizados** que possibilitem a avaliação comparativa de abordagens explicativas guiadas por semântica estruturada, ampliar a investigação sobre **integração entre pipelines subsimbólicos e estruturas simbólicas formais por meio da IA neurossimbólica**, com vistas a garantir maior robustez estatística e inteligibilidade conceitual.

A consolidação da XAI como uma disciplina verdadeiramente transparente, confiável e centrada no ser humano dependerá da combinação entre avanços técnicos e fundamentos semânticos sólidos, possibilitados pelo uso criterioso de ontologias bem estruturadas.

Declaration on Generative AI

In the preparation of this manuscript, the authors made limited use of the ChatGPT-4o generative AI model. The tool was employed exclusively for linguistic assistance, including rephrasing, improving readability and style, and checking grammar and spelling, as outlined in the CEUR-WS GenAI Usage Taxonomy. No part of the research design, analysis, or interpretation of results was carried out by AI. All suggested edits were critically reviewed, adapted, and validated by the authors, who take full responsibility for the final content of this publication.

Referências

[1] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Bmj (2021) n71. doi:10.1136/bmj.n71.

[2] H. Sheikh, C. Prins, E. Schrijvers, Artificial Intelligence: Definition and Background, Springer International Publishing, Cham, 2023, pp. 15–41. doi:10.1007/978-3-031-21448-6_2.

[3] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[4] R. Studer, V. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data & Knowledge Engineering 25 (1998) 161–197. doi:10.1016/s0169-023x(97)00056-6.

[5] W. Maass, V. C. Storey, Pairing conceptual modeling with machine learning, Data amp; Knowledge Engineering 134 (2021) 101909. doi:10.1016/j.datak.2021.101909.

[6] S. Ali, T. Abuhmed, S. El-Sappagh, et al., Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information Fusion 99 (2023) 101805. doi:https://doi.org/10.1016/j.inffus.2023.101805.

[7] E. Rajabi, K. Etminani, Knowledge-graph-based explainable ai: A systematic review, Journal of Information Science 50 (2022) 1019–1029. URL: http://dx.doi.org/10.1177/01655515221112844. doi:10.1177/01655515221112844.

[8] R. Confalonieri, L. Coba, B. Wagner, et al., A historical perspective of explainable Artificial Intelligence, WIREs Data Mining and Knowledge Discovery 11 (2021) e1391. doi:10.1002/widm.1391.

[9] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Fat* '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 279–288. doi:10.1145/3287560.3287574.

[10] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.

[11] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, 2017. doi:10.48550/arxiv.1705.07874.

[12] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Mining and Knowledge Discovery 38 (2024) 3043–3101. doi:10.1007/s10618-022-00867-8.

[13] A. d. Garcez, L. C. Lamb, Neurosymbolic AI: the 3rd wave, Artificial Intelligence Review 56 (2023) 12387–12406. doi:10.1007/s10462-023-10448-w.

[14] S. Badreddine, A. d'Avila Garcez, L. Serafini, et al., Logic Tensor Networks, Artificial Intelligence 303 (2022) 103649. doi:10.1016/j.artint.2021.103649.

[15] G. Guizzardi, A. Botti Benevides, C. M. Fonseca, et al., UFO: Unified Foundational Ontology, Applied Ontology 17 (2022) 167–210. doi:10.3233/ao-210256.

[16] S. Borgo, R. Ferrario, A. Gangemi, et al., Dolce: A descriptive ontology for linguistic and cognitive engineering, Applied Ontology 17 (2022) 45–69. doi:10.3233/ao-210259.

[17] J. N. Otte, J. Beverley, A. Ruttenberg, Bfo: Basic formal ontology, Applied Ontology 17 (2022) 17–43. doi:10.3233/ao-220262.

[18] G. Guizzardi, N. Guarino, Explanation, semantics, and ontology, Data Knowledge Engineering 153 (2024) 102325. doi:https://doi.org/10.1016/j.datak.2024.102325.

[19] R. Confalonieri, G. Guizzardi, On the multiple roles of ontologies in explainable ai, 2023. doi:10.48550/arxiv.2311.04778.

[20] S. Chari, D. M. Gruen, O. Seneviratne, et al., Directions for explainable knowledge-enabled systems, 2020. doi:10.48550/arxiv.2003.07523.

[21] P. Sun, Y. Wu, C. Yin, et al., Molecular Subtyping Of Cancer Based On Distinguishing Co-expression Modules And Machine Learning, Front Genet 13 (2022). doi:10.3389/fgene.2022.866005.

[22] C. Lu, S. Pathak, G. Englebienne, et al., Channel Contribution In Deep Learning Based Automatic Sleep Scoring - How Many Channels Do We Need?, Ieee Trans Neural Syst Rehabil Eng 31 (2023). doi:10.1109/tnsre.2022.3227040.

[23] Y. Li, Y. Liu, J. Hou, et al., A Center-anchored Adaptive Hierarchical Graph Neural Network With Application In Structure-aware Recognition Of Enzyme Catalytic Specificity, Neurocomputing 619 (2025). doi:10.1016/j.neucom.2024.129155.

[24] Z. Liu, Z. Zhang, X. Zeng, et al., Representation And Association Of Chinese Financial Equity Knowledge Driven By Multilayer Ontology, Data Inf Manag 6 (2022) 3.0. doi:10.1016/j.dim.2022.100009.

[25] O. Perera, J. Liu, Exploring Large Language Models For Ontology Learning, Issue Inf Syst 25 (2024) 4.0. doi:10.48009/4_iis_2024_124.

[26] X. Ma, T. Wu, G. Li, et al., Dse-hngcn: Predicting the frequencies of drug-side effects based on heterogeneous networks with mining interactions between drugs and side effects, Journal of Molecular Biology 437 (2025) 168916. doi:https://doi.org/10.1016/j.jmb.2024.168916, emerging Artificial Intelligence Methodologies in Computational Biology.

[27] Y. Wei, T. Zhang, B. Wang, et al., Indelpred: Improving The Prediction And Interpretation Of Indel Pathogenicity Within The Clinical Genome, Hum Genet Genom Adv 5 (2024) 4.0. doi:10.1016/j.xhgg.2024.100325.

[28] S. Zhao, Z. Cui, G. Zhang, et al., Mgppi: Multiscale Graph Neural Networks For Explainable Protein–protein Interaction Prediction, Front Genet 15 (2024). doi:10.3389/fgene.2024.1440448.

[29] L. Rasbach, A. Caliskan, F. Saderi, et al., An Orchestra Of Machine Learning Methods Reveals Landmarks In Single-cell Data Exemplified With Aging Fibroblasts, Plos One 19 (2024) 4.0. doi:10.1371/journal.pone.0302045.

[30] S. Dubey, S. Singh, D. Verma, et al., Genocare Prognosticator Model: Host Genetics Predict Severity Of Infectious Disease, Scalable Comput Pract Exp 26 (2025) 2.0. doi:10.12694/scpe.v26i2.3774.

[31] Y. Zhao, J. Shao, Y. Asmann, Assessment And Optimization Of Explainable Machine Learning Models Applied To Transcriptomic Data, Genomics Proteomics Bioinform 20 (2022) 5.0. doi:10.1016/j.gpb.2022.07.003.

[32] C. Zhan, S. Ren, Y. Zhang, et al., Mio: An Ontology For Annotating And Integrating Medical Knowledge In Myocardial Infarction To Enhance Clinical Decision Making, Comput Biol Med 190 (2025). doi:10.1016/j.compbiomed.2025.110107.

[33] I. Sung, S. Lee, D. Bang, et al., Mdtr: A Knowledge-guided Interpretable Representation For Quantifying Liver Toxicity At Transcriptomic Level, Front Pharmacol 15 (2024). doi:10.3389/fphar.2024.1398370.

[34] S. Li, F. Dong, R. Li, et al., A Dual Medical Ontology And Relational Graph Framework With Neural Ordinary Differential Equations For Diagnostic Prediction, Neurocomputing 647 (2025). doi:10.1016/j.neucom.2025.130519.

[35] X. Zheng, D. Meng, D. Chen, et al., Sccat: An Explainable Capsulating Architecture For Sepsis Diagnosis Transferring From Single-cell Rna Sequencing, Plos Comput Biol 20 (2024) 10.0. doi:10.1371/journal.pcbi.1012083.

[36] Z. Chen, Z. He, X. Liu, et al., Evaluating Semantic Relations In Neural Word Embeddings With

Biomedical And General Domain Knowledge Bases, Bmc Med Informatics Decis Mak 18 (2018). doi:10.1186/s12911-018-0630-x.

[37] A. Canito, J. Corchado, G. Marreiros, A Systematic Review On Time-constrained Ontology Evolution In Predictive Maintenance, Artif Inl Rev 55 (2022) 4.0. doi:10.1007/s10462-021-10079-z.

[38] K. Meghraoui, T. Racharak, K. Ait, et al., A New Integrated Neurosymbolic Approach For Crop-yield Prediction Using Environmental Data And Salite Imagery At Field Scale, Artif Inl Geosci 6 (2025) 1.0. doi:10.1016/j.aiig.2025.100125.

[39] C. Guan, A. Gong, Y. Zhao, et al., Interpretable Machine Learning Model For New-onset Atrial Fibrillation Prediction In Critically Ill Patients: A Multi-center Study, Crit Care 28 (2024) 1.0. doi:10.1186/s13054-024-05138-0.

[40] A. Basakin, Y. Kulchin, V. Gribova, et al., Prospects For The Development Of Laser-based Directed Energy Deposition Additive Process Based On Ai Technologies, Bull Russ Acad Sci Phys 88 (2024) 3.0. doi:10.1134/s1062873824709711.

[41] Q. Li, Y. Yu, P. Kossinna, et al., Xa4c: Explainable Representation Learning Via Autoencoders Revealing Critical Genes, Plos Comput Biol 19 (2023) 10.0. doi:10.1371/journal.pcbi.1011476.

[42] F. Kaptein, B. Kiefer, A. Cully, et al., A Cloud-based Robot System For Long-term Interaction: Principles, Implementation, Lessons Learned, Acm Trans Hum Robot Interact 11 (2022) 1.0. doi:10.1145/3481585.

[43] N. Smedley, D. Aberle, W. Hsu, Using Deep Neural Networks And Interpretability Methods To Identify Gene Expression Patterns That Predict Radiomic Features And Histology In Non-small Cell Lung Cancer, J Med Imaging 8 (2021) 3.0. doi:10.1117/1.jmi.8.3.031906.

[44] C. Ruiz-Arenas, I. Mar, L. Wang, et al., Netactivity Enhances Transcriptional Signals By Combining Gene Expression Into Robust Gene Set Activity Scores Through Interpretable Autoencoders, Nucleic Acids Res 52 (2024) 9.0. doi:10.1093/nar/gkae197.

[45] L. Seninge, I. Anastopoulos, H. Ding, et al., Vega Is An Interpretable Generative Model For Inferring Biological Network Activity In Single-cell Transcriptomics, Nat Commun 12 (2021) 1.0. doi:10.1038/s41467-021-26017-0.

[46] M. Singha, L. Pu, G. Srivastava, et al., Unlocking The Potential Of Kinase Targets In Cancer: Insights From Canceromicsnet, An Ai-driven Approach To Drug Response Prediction In Cancer, Cancers 15 (2023) 16.0. doi:10.3390/cancers15164050.

[47] K. Askland, D. Strong, M. Wright, et al., The Translational Machine: A Novel Machine-learning Approach To Illuminate Complex Genetic Architectures, Genet Epidemiol 45 (2021) 5.0. doi:10.1002/gepi.22383.

[48] H. Townsend, K. Rosenberger, L. Vanderlinden, et al., Evaluating Methods For Integrating Single-cell Data And Genetics To Understand Inflammatory Disease Complexity, Front Immunol 15 (2024). doi:10.3389/fimmu.2024.1454263.

[49] G. A. Miller, WordNet: A lexical database for English, in: Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994, 1994.

[50] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Research 32 (2004) 267d–270. doi:10.1093/nar/gkh061.

[51] E. G. Brown, L. Wood, S. Wood, The medical dictionary for regulatory activities (meddra), Drug Safety 20 (1999) 109–117. doi:10.2165/00002018-199920020-00002.

[52] M. Ashburner, C. A. Ball, J. A. Blake, et al., Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. doi:10.1038/75556.

[53] Y. Li, Y. Liu, J. Hou, et al., A center-anchored adaptive hierarchical graph neural network with application in structure-aware recognition of enzyme catalytic specificity, Neurocomputing 619 (2025) 129155. doi:https://doi.org/10.1016/j.neucom.2024.129155.

[54] N. Brandes, D. Ofer, Y. Peleg, et al., Proteinbert: a universal deep-learning model of protein sequence and function, Bioinformatics 38 (2022) 2102–2110. doi:10.1093/bioinformatics/btac020.

[55] W. Maass, V. C. Storey, Pairing conceptual modeling with machine learning, Data Knowl. Eng. 134 (2021) 101909. doi:10.1016/j.datak.2021.101909.

[56] G. Amaral, F. Baião, G. Guizzardi, Foundational ontologies, ontology-driven conceptual modeling,

and their multiple benefits to data mining, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11 (2021) e1408.

Improving Semantic Expressiveness in Data Catalogs: Applying Ontological Patterns to Data Catalog Vocabulary Relations

Vânia Borges[1,*], Natália Queiroz de Oliveira[1], Maria Luiza Machado Campos[1], and Giseli Rabello Lopes[1]

[1] *Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil*

Abstract

On the Web, cataloged resources can be related in many ways. Complex relationships may be necessary to characterize the context in which the resources were created, allowing for tracing input data, the software used, and the agents and funders involved. However, describing these relationships expressively and formally, contributing to the semantics of the resources associated, and facilitating interoperability is still a challenge. To avoid inconsistencies and ambiguities between different interpretations of relations, it is essential to use common terminology. The Data Catalog Vocabulary (DCAT) is a W3C-recommended schema used for catalog interoperability, modeling the data structures of relevant resources along with their primary relationships. This work proposes applying ontological patterns based on the Unified Foundational Ontology (UFO) to improve the semantic expressiveness of domain-specific relationships that are not currently offered in DCAT. The use of these patterns offers additional details about interactions among those involved and, when needed, documents the evolution of relationships over time, contributing to understanding, reuse, and interoperability.

Keywords

Data Catalog Vocabulary, Relations, Relationships, Ontological Patterns.

1. Introduction

Data catalogs have been gaining prominence in the literature as a solution for increasing visibility and access to cataloged resources [1,2]. They are collections of metadata, combined with data management and search tools, that assist analysts and other data users in finding the data they need [3]. Serving as an inventory of available resources, they provide information to evaluate the fitness of data for intended uses [4]. Their metadata are organized into schemas, also known as metadata models, which capture information about various aspects of a cataloged resource [5], including how they relate to each other. These cataloged resources often originate from diverse data repositories, encompassing a wide array of datasets that require proper description and discoverability. Complex relationships may be necessary to describe the context in which these resources were created, enabling tracing of input data, the software used, and the agents and funders involved [6]. However, describing these relationships expressively and formally, contributing to the semantics of the associated resources, and facilitating interoperability remains a challenge.

In the catalog domain, Data Catalog Vocabulary (DCAT) is a W3C recommendation for catalog interoperability [7]. It has been used in different implementations such as: the DCAT Application Profile (DCAT-AP) [8] that serves as a standard for describing public sector datasets across Europe; the GeoDCAT-AP that represents geographic metadata in European data portals [9]; and the core

[1] *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.*

[*] Corresponding author.

vjborges30@ufrj.br (V. Borges); natalia.oliveira@ppgi.com.br (N. Q. Oliveira); mluiza@ppgi.com.br (M. L. M. Campos); giseli@ic.ufrj.br (G. R. Lopes)

0000-0002-6717-1168 (V Borges); 0000-0001-8371-142X (N. Q. Oliveira); 0000-0002-7930-612X (M. L. M. Campos); 0000-0003-2482-1826 (G. R. Lopes)

of the FAIR Data Point schema, which facilitates the publication of FAIR-compliant metadata for catalogued resources [10]. DCAT describes cataloged resources on the Web with an emphasis on datasets and data services. Thus, a publisher can describe their datasets and data services in a catalog using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from various catalogs [7]. DCAT offers a set of relations between resources and proposes solutions for representing domain-specific relations, which include the use of *dcat:Relationship*. According to DCAT, *dcat:Relationship* applies to a specific set of relationships, i.e., those in which one resource plays a role with respect to another. This type of relation, known as role-playing relation, is common in social or organizational situations, where the dynamic between the parties is mediated by specific roles that each party plays [11]. This work focuses on using ontological patterns to enhance the expressiveness and formalism of these relations, providing mechanisms for catalogs to support the dynamic creation of these relationships at the schema level and, as a result, for reuse by resource publishers.

Patterns are instruments for encapsulating common knowledge that can contribute to the analysis of different concepts and types of relations, thereby improving representation and providing support for machine-actionability for catalogs. In the Software Engineering community, the term “pattern language” refers to a network of interrelated patterns together with a process for systematically solving coarse-grained software development problems [12]. This approach has been successfully applied in ontology engineering through the development of ontology patterns (OPs). OPs are an emerging approach that benefits the reuse of encoded experiences and good practices [13], giving rise to ontology pattern languages (OPLs). The literature highlights different contexts that explore OPs to enhance semantic expressiveness. For instance, in multidimensional models used in the representation of analytical data in data warehouses [14]; in OWL ontologies, with an alignment between the I-ADOPT framework and the OPs established by Measurement OPL (M-OPL) [15]; and in guiding the definition of exploratory questions for conceptual models, guaranteeing pragmatic explanations in relation to the model [16].

Recent work has presented a systematic analysis of truthmaking patterns (TMP) for relations, based on the ontological nature of their truthmakers, which are the entities responsible for the truth of the propositions arising from the relationships [11]. It presents several TMPs for relations with different levels of expressivity. As ontological patterns, these TMPs help define concepts and relationships, speeding up ontology development and encouraging reuse.

This work proposes the adoption of the TMP and powertype pattern for improving the semantics of the role-playing relations in DCAT. The goal is to enhance the semantics of DCAT by using TMP for descriptive relationships. Instead of relying on relationships with embedded semantics in the data, a pattern is suggested to guide the catalog administrators in creating domain-specific relations and their truthmakers (relationships). These relations, shown in the schema, become reusable by publishers. The truthmakers provide details about interactions among those involved and, when needed, document how these relationships evolve over time. Adding this information helps agents better understand relationships and improves interoperability.

This paper is organized as follows: Section 2 presents the background of the paper. Section 3 describes how DCAT handles relations and relationships. Section 4 applies the truthmaking pattern to *dcat:Relationship*. Section 5 addresses a simple implementation in OWL to explore the potential of the TMP. Finally, in Section 6, we conclude and list some future work.

2. Ontological patterns for relations in UFO

The terms "*relation*" and "*relationship*" are often used interchangeably in the literature. This paper considers the distinction proposed by Guarino and Guizzardi [17]. According to the authors, a *relation* holds because the *relationship* exists. In this case, the authors identify the relationship as the truthmaker of the relation, i.e., what establishes the truth of the propositions derived from this relationship.

The UFO relation taxonomy utilizes two orthogonal distinctions of relations that consider internal/external and descriptive/non-descriptive relations [11, 18]. Given the objective of this paper, we focus here on external-descriptive relations. In UFO, material relations are external descriptive relations that hold in virtue of at least one *relational moment* inhering in at least one *relatum* that is existentially dependent on another *relatum.* They can be single-sided whenever they hold in virtue of one or more moments inhering in just one *relatum* (e.g., <administratorOf>); or multi-sided whenever they hold in virtue of at least two moments, each inhering in a different *relatum* (e.g., <treatedIn> between a patient and a medical unit) [11].

Guarino et al. [11] propose a systematic analysis of truthmaking patterns for relations, based on the ontological nature of their truthmakers. In their work, they present a number of TMP for relations according to levels of expressivity. Before proceeding, it is essential to discuss an important notion, namely, the distinction between strong and weak truthmakers. A strong truthmaker is one whose existence is sufficient for a proposition to be true. In contrast, a weak truthmaker makes a proposition true not merely because of its existence, but because of the way it contingently is. In this paper, we focus on the TMP for external descriptive relationships, especially role-playing relations [11]. In this type of relation, an entity plays a relational role that emerges due to the existence of another entity. To illustrate, Figure 1 depicts the TMPs for this kind of descriptive relation, as represented in Figure 1(a). A weak TMP is illustrated in Figure 1(b), where the material relation <administratorOf> between an admin and a catalog is derived from the relator <Administration>. This relator accounts for the social commitments and obligations associated with the administrative role <Admin>, which depends on some catalog. Figure 1(c) shows the full TMP, which adds the event <AdministrationEvolution>. The event is a strong truthmaker and accounts for the period of time during which the role is played.

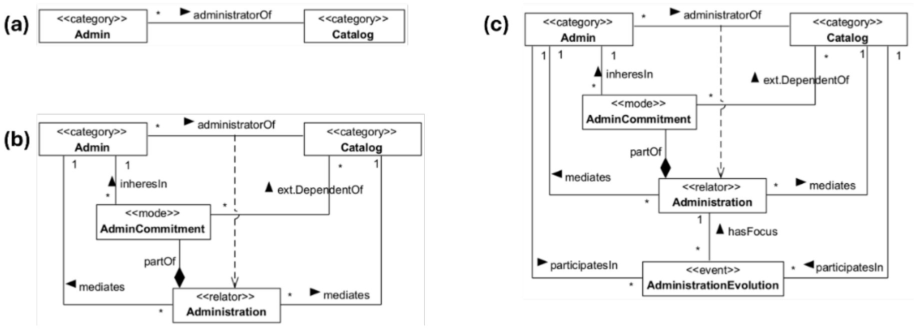


Figure 1: Weak and full truthmaking patterns for a single-sided relation.

The powertype pattern is a well-known pattern in the conceptual modeling community and is relevant to this research. This pattern addresses a phenomenon that occurs in various subject domains that require the handling of multiple classification levels [14, 19]. It is an example of an early approach for multi-level modeling in software engineering, used to model situations in which instances of a type (the power type) are specializations of a lower-level type (the base type), and both power types and base types appear as regular classes in the model. Multi-Level Theory (MLT) [20] is an axiomatic theory based on UFO that provides powertype support in UML [19]. It uses the <<instantiation>> stereotype to highlight the association between the base type and the higher-order type, establishing an instantiation semantics. According to the theory, the cardinality between the types involved establishes distinct behaviors for instances of the base type.

Figure 2 shows an example where each instance of <Agent> will be an instance of, at most, one instance of the higher-order type, in this case, an instance of <Person> or <Organization>. For a complete description of the approach, see [19].

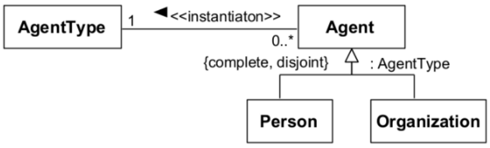


Figure 2: Relation between higher-order type and base type.

3. Relations and relationships in DCAT

DCAT is a metadata vocabulary implemented in OWL 2 that reuses terms from standardized vocabularies, such as Dublin Core (DC), Friend Of A Friend (FOAF), and Provenance Ontology (PROV-O). Additionally, it defines a minimal set of classes and properties of its own [21]. Figure 3 shows a UML diagram of the DCAT entities associated with the definition of qualified relationships. The model adopts the following prefixes: *dcat* for DCAT concepts, *foaf* for Friend Of A Friend vocabulary, *skos* for Simple Knowledge Organization System, and *dct* for Dublin Core Terms. In the figure, *dcat:Resource* represents the cataloged resources, i.e., resources published or curated by a single agent. According to DCAT, this class should not be instantiated, but rather its specializations.

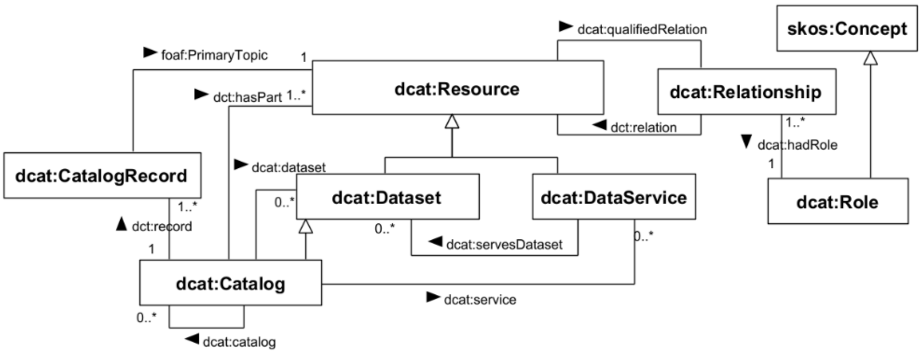


Figure 3: Simplified view of DCAT resources and qualified relationships.

In addition to the relations defined between entities, DCAT also includes classes to facilitate the creation of new ones. This flexibility is necessary because resources can be related in many ways. Furthermore, complex relationships may be needed to characterize the context in which the resources were created, for example, by tracing input data, the software used, and the agents and funders (sponsors/financiers) involved [6].

DCAT offers qualified relations to support complex non-binary relations not covered by PROV-O and DCAT properties. These qualified relations can also be convenient when relations are represented using known properties but have additional information needs that require a more sophisticated representation [6]. For example, one might want to describe the temporal dimension of a function, i.e., the period during which an individual or organization performed a certain function.

To create relations across resources, DCAT utilizes the *dcat:qualifiedRelations* property [6]. This property links the source resource to an instance of the *dcat:Relationship.* In the DCAT specification, *dcat:Relationship* is "an association class for attaching additional information to a relationship between DCAT Resources" [21]. It involves another resource referenced by the *dct:relation* property, which, in the context of a *dcat:Relationship*, must point to another *dcat:Dataset* or another cataloged resource [6]. The *dcat:Relationship* uses the *dcat:hadRole* property to link to *dcat:Role.* The *dcat:Role* class has two functions in the specification. It provides the meaning of the agent responsibility regarding the Resource and the role or function of an Entity concerning another provided Entity [6].

According to the current schema, new relationships are represented along with the instances (data), making it difficult to standardize and reuse relations in the model. According to Albertoni *et al.* [21], these expected relationships can be complex and must be addressed. For DCAT, associating a term from a semantic artifact is sufficient to provide the semantics of these relationships. However, it does not establish the nature of the relationship or provide a means of understanding it. The following sections explore the TMP for descriptive relations and the powertype pattern to assist in the implementation of new relations.

4. Applying the truthmaking pattern for descriptive relations

The *dcat:Relationship* entity was introduced into DCAT from its second version onwards to enable the representation of relationships between datasets and other resources. According to the DCAT, it applies to a specific set of relations, i.e., those in which one resource plays a role with respect to another. As aforementioned, this type of relation, known as role-playing relation, is common in social or organizational situations, where the dynamic between the parties is mediated by specific roles that each party plays [11].

One approach to representing role-playing relations involves using relators [11]. In this approach, the relator expresses the social commitments and obligations associated with the roles that individuals or entities play in a given context. Reification of the relator, as the truthmaker of these relationships, provides a more explicit structure for understanding the interactions. Additionally, explicitly defining the roles played by entities helps establish the cardinalities of the relationships, thereby avoiding ambiguities in conceptual modeling. Another important aspect is the dynamism that can occur in this type of relationship over time, depending on the circumstances and actions of the parties involved. In this case, as presented in the full TMP, events can be relevant to documenting the properties evolution of these relationships [11].

Based on the definitions of relation and relationship mentioned in Section 2, we consider that entities under the superclass *dcat:Relationship* are those whose instances serve as truthmakers for relations between resources. These entities are categorized as relators in UFO, corresponding to the mereological sum of external dependent aspects of at least one entity involved. As an association class[2], it also allows modelers to define relationship-specific properties. As a superclass, *dcat:Relationship* establishes mandatory properties for distinct relationship types in the domain. Accordingly, it is classified as a UFO category.

To standardize and define new relationships and relations across resource types at the schema level, it is essential that the value associated with *dcat:Role* be explicitly expressed in the model rather than alongside the data. In this way, entities that specialize in *dcat:Relationship* become carriers of this relational aspect that refers to one of the entities involved in the relationship. To achieve this, it is necessary to deal with *dcat:Role* as a higher-order type. It is worth noting that UFO distinguishes between first-order types (1stOTs) and high-order types (highOTs) based on their level of abstraction and categorization of entities represented. While 1stOTs represent individual concrete objects, highOTs represent categories of 1stOTs or other highOTs.

4.1. dcat:Role as high-order type

Based on the DCAT specification, *dcat:Role* instances are terms in a semantic artifact that specify the meaning of an entity role concerning another or an agent role regarding an entity. This definition corresponds to "the position or purpose that someone or something has in a situation, organization, society, or relationship" as stated in the Cambridge Dictionary[3]. In this context, it represents a relevant aspect of an entity that is externally dependent on another. Therefore, we consider *dcat:Role* an external dependent aspect representing a relational property of an agent or source resource that depends on another resource. External dependent aspects define the properties that an individual holds in the scope of a certain material relation [23].

Addressing semantic overload and distinguishing the ontological nature of *dcat:Role*, this paper considers it as a superclass to denote distinct roles and functions based on their application: *ResourceRole* and *AgentRole*. Each of these new quality types is linked to its own value space. These value spaces – referred to as quality structures - are abstract entities delimiting the range of possible values (qualia, singular quale) for qualities of a given quality type [24]. Therefore, each specialization of *dcat:Role* defines a conceptual space that must be managed by a nominal quality structure [25]. Each potential value can be a term with an independent meaning. Many semantic

[2] An Association that has a set of Features of its own [22].

[3] https://dictionary.cambridge.org/us/dictionary/english/role

artifacts, such as thesauri and taxonomies, have structural relations that software agents can utilize to identify synonyms and equivalents with other artifacts.

4.2. A pattern for relationship types

Based on the distinction between 1stOTs and highOTs, we propose a pattern for role-playing relations in DCAT using TMP, *dcat:Role* and *dcat:Relationship*. It should be noted that the use of entities such as *dcat:Role* to link terms that indicate the meaning of the role or function performed by another entity is not exclusive to DCAT. Other vocabularies, such as Bibliographic Framework Initiative[4] (BIBFRAME) and Organization Ontology[5] (ORG), use classes in similar ways.

Our pattern outlines the relations between *relata*, where at least one aspect of a *relatum* depends externally on another. Consequently, it permits the specification of a relationship between two or more resources where an aspect of one resource (*ResourceRole*) is externally dependent on another resource(s). The full TMP of interest comprises a relator and an event. The relator connects the involved entities (related resource types) and adds characteristics of the *relata* that are externally dependent on another, in this case, the role of the resource. This relator may also possess its own characteristics. One or more material relations between the entities involved will be derived from the instantiated relator. This way, the relator makes the semantics of the material relation explicit.

Here, we propose that entities aligned with the pattern are modeled as highOTs, aiding catalog modelers in defining new relationships and relations. These entities are shown in Figure 4. The models presented in this paper were implemented using the Visual Paradigm[6] tool version 17.2. This tool has a plugin[7] for OntoUML, an ontology-driven modeling language that incorporates the distinctions underlying UFO into UML class diagrams [24]. It introduces various stereotypes that correspond to the concepts defined in UFO, as well as grammatical formal constraints that reflect its axiomatization.

Using the powertype pattern, the cataloged resource type (*ResourceType*) is powertype of *dcat:Resource* entity. As a result, all the specializations of this entity become instances of the *ResourceType*, and instances of *dcat:Resource* must be an instance of, at most, one instance of resource type. Related resource type (*RelatedResourceType*) and resource type by role (*ResourceTypeByRole)* are specializations of *ResourceType. ResourceTypeByRole* identifies domain resource types according to their role and categorizes *dcat:Resource*. In this context, its instances are a set of *dcat:Resource* specializations that assume a role in relation to other resources. The *RelatedResourceType* classifies specific resource types in the domain that are involved in relationships. When made explicit at the schema level, the instances of these highOTs are anti-rigid types whose contingent classification condition is relational. Therefore, they are categorized as <Rolemixin>.

RelationshipType is the powertype of *dcat:Relationship*. It is rigid and categorized as <Category>. *RelationshipTypeByRole* specializes *RelationshipType* and categorizes *dcat:Relationship*. Its instances are relators that mediate resource types and are truthmakers for role-playing relations between them. In defining relationships, *RelationshipTypeByRole* is a sortal, and its instances must be types classified as <Relator> that corresponds to the mereological sum of the external dependent moment of at least one resource type, including the *ResourceRole* instance.

[4] https://www.loc.gov/bibframe/
[5] https://www.w3.org/TR/vocab-org/
[6] https://www.visual-paradigm.com/
[7] https://github.com/OntoUML/ontouml-vp-plugin

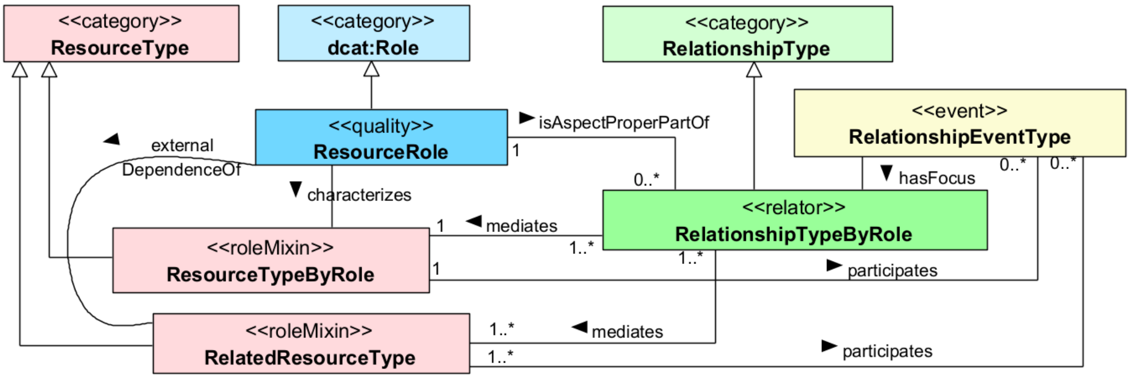


Figure 4: Full pattern for relationships among resource types in ML-DCAT.

The full TMP also employs events. For the DCAT relationships, the events can work as relational episodes [17]. In this context, focusing on the relator, an event allows us to follow the manifestation of certain aspects defined in the relationship [17]. Instances of its instances can, for example, follow the time period during which a function or role is performed. Thus, the *RelationshipTypeEvent* entity is defined as an <Event Type> that focuses on *RelationshipTypeByRole*. Consequently, it captures the creation of relationship types and the manifestation of specific properties (aspects) of those relationships at a given time. The way the episode reports participation varies based on the relationship specific properties. According to Guarino, Sales and Guizzardi [11], these aspects are considered the "focus" of the event, while the relationship itself is the focus of the relational episode. It is through the relationship that the event is understood and interpreted. Instances of *RelationshipEventType* are created only when changes in types are significant to the domain. They allow temporal and contextual metadata to be associated with the relationship, such as property changes over time in a domain relationship, as well as the responsible agent. This work has adopted behavior similar to that of highOT endurants for high-order events, including the use of intra- and cross-level structural relations [20]. Thus, the model incorporates *RelationshipEvent* as the base type of the *RelationshipEventType* powertype. Similar to other DCAT entities, its instance depends on the level of change monitoring that the catalog intends to employ.

The *ResourceRole* entity (*dcat:Role* specialization) is an external dependent moment, an aspect, within the pattern. It characterizes *ResourceTypeByRole* and depends on *RelatedResourceType*, both entities mediated by *RelationshipTypeByRole*, which formalizes and makes explicit this dependency as a mereological sum of externally dependent aspects. According to multi-level theory, the entity *ResourceRole* is a ***regularity property***, i.e., it is an attribute defined in a highOT that influences the intension of instances of this type [20]. Aligned with this vision, *ResourceRole* fulfills the DCAT proposal, establishing the meaning of the relationship, and contributes to the intension of the material relations dynamically defined at the schema level. In the *RelationshipTypeByRole* instances, the *ResourceRole* instances are standardized values listed in Quality Structures. They capture the role of the resource regarding another, offering the meaning for relationships and relations that will be instantiated. The modeler can also insert relevant characteristics about the resources that influence the relationship, as well as those specific to the relationship itself. These characteristics can also change over time, depending on the contextual circumstances of the relationship.

Adopting the pattern provides the modeler with additional resources to analyze the addressed concepts, offering tools to improve their representation. The pattern also allows the modeler to define grounded role-playing relations. Furthermore, it is possible to observe an evolution of the concept associated with the *dcat:Relationship* class. This class is no longer just a UML association class and is now treated as a relator. While the former is identical to an instance of an association, i.e., an objectified n-tuple, the relator grounds those n-tuples, in the sense that its instances represent what happens in reality whenever the association holds [26]. Associations, in turn, are represented by material relations derived from relators.

To define new material relations, the following steps must be taken: (i) identify the types of resources to be connected; (ii) choose a term from a standardized vocabulary that accurately represents the semantics of the resource role; (iii) clarify (made explicit) the resource role as a new type, if needed; (iv) model the relationship by specializing *dcat:Relationship*; (v) add valued attributes into the object facet of the relationship instance; (vi) define additional properties to the class facet, if needed; (vii) associate the involved resources; (viii) link the resource types, defining one or more material relations; and (ix) evaluate whether an event that works as a mechanism for recording and monitoring the evolution of the relationship over time is needed. These relations must be linked to the relator from which they derive. It is important to note that external dependence relations and their reified relators address cardinality issues and the ambiguity between association specialization, subsetting and redefinition [27, 28].

5. Employing Well-founded Relations among Resources

To illustrate the use of the pattern for role-playing relations, we present part of a model developed for a healthcare catalog that describes research datasets on patient clinical data and its operational ontology. These datasets were generated from electronic medical records and processed to fit into a research case record form. To ensure provenance, metadata from the workflow applied in the process was created and packaged in a complementary dataset. Similarly, metadata from the extraction, processing, and conversion to the form was also produced and exported to another dataset. These datasets contain, respectively, the prospective and retrospective provenance metadata of the clinical data dataset. To publish these datasets and the relationships between them, the catalog uses DCAT and the pattern presented in Section 4.

In this section, we present the conceptual model implemented to represent the relationship and its transformation to an OWL ontology, which functions as a well-founded semantic model for the catalog. This semantic model is published in the same triplestore where the dataset descriptors are published, providing different agents with relevant information about the model and the existing data. For this implementation, we used the Systematic Approach for Building Ontologies (SABiO), a methodology that enables the development of operational ontologies from a reference ontology — a conceptual model that clearly and accurately describes the entities in the domain [29]. Following this methodology, the conceptual model is developed using Visual Paradigm and the OntoUML plugin. With the plugin, the conceptual model is translated into a structure defined by gentle UFO (gUFO), a lightweight version of UFO implemented in OWL 2 to support the design of well-founded operational ontologies [30]. It is then exported as a serialized OWL file in Turtle (TTL). To assist modelers, two applications implemented in Jupyter Notebook are used to adjust the ontology. These applications will be discussed in a future paper. The ontology is imported into Protégé[8] tool, where adjustments and validations with plugins and reasoners are performed. Once completed, it is published in the GraphDB[9], a triplestore used to store (meta)data in triples.

5.1. Conceptually modeling new relationships

Figure 5 illustrates a UML model with a relationship where one dataset works as the provenance dataset for the others, meaning its data represents the provenance metadata for the related datasets. According to the figure, a dataset can have more than one dataset with provenance, covering prospective and retrospective provenance metadata. On the other hand, a dataset can contain provenance metadata from more than one dataset. The highOTs are represented with the stereotype <<type>> for enhanced visibility. In the model, DCAT classes retain the prefix *dcat*; classes referring to the relation pattern are associated with the prefix *mldcat*, in reference to multi-level entities for DCAT. Domain-specific relationships use the prefix *mycat*. The prefix *datacite_voc* refers to the Datacite[10] vocabulary. The provenance dataset is explicitly defined as a specialization

[8] https://protege.stanford.edu/

[9] https://graphdb.ontotext.com/

[10] http://purl.org/datacite/v4.4/

of the dataset. *mycat:ProvenanceRelationship* represents the *mldcat:RelationshipTypeByRole* instance, with *datacite_voc:isMetadataFor* outlining the relationship intension. The term means that the data in a dataset is metadata for another resource, serving to represent the semantics of the *mycat:ProvenanceDataset*. The *mycat:isProvenanceMetadataFor* is a material relation derived from this relator. Based on the pattern, the term refers to the provenance dataset (intrinsic aspect) and is externally dependent on the existence of the dataset to which it relates. Attributes can be defined in the relator. Therefore, the relationship can indicate that *dataset* A partially covers the provenance of *dataset* B and fully covers that of *dataset* C, while also recording the level of confidence in that provenance.

In the model, the event for documenting changes adds the agents involved. It enables tracking changes to provenance metadata about the datasets, documenting updates to the catalog, such as changes to the confidence level of the provenance and the agent responsible for them. Similarly, the same treatment can be offered to improve agent attributions regarding resources in DCAT. This topic has not been explored here due to the limited number of pages.

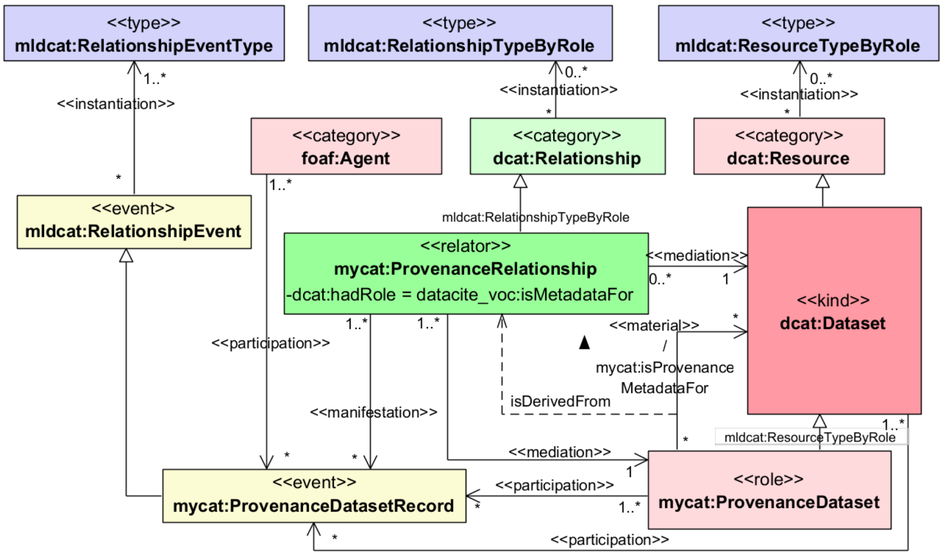


Figure 5: Example of Material Relation across Datasets using Truthmaking Pattern.

The pattern also applies to other domains, such as portals like OASIS BR[11] — a Brazilian portal that brings together scientific production and research data in open access, published on different digital platforms. In this case, it can define the relationship between the aggregator catalog (Portal) and the platforms it refers to. This relationship can include attributes such as periodicity and harvest type, which may change over time, such as switching the harvest periodicity from monthly to fortnightly. In this case, the event serves as a record mechanism to track these changes.

It should be noted that accidental or contingent roles played by types enhance the model semantics when made explicit. These types can be suppressed in the design phase to ensure the metadata schema fulfills non-functional requirements. To aid this process, UFO offers a method that establishes rules for abstraction [31].

5.2. Publishing the Semantic Model and Descriptors in the Catalog Triplestore

After validations made in Protégé, the ontology file (semantic model) is imported into the GraphDB triplestore along with the descriptors (cataloged resources), instances of the model. Figure 6 presents a simplified graph view of the relationship and material relation illustrated in Figure 5, published in a GraphDB triplestore. The categories of a foundational ontology, such as UFO, establish a common language and a referential model that can be used to describe the schema types and their relations [32]. Based on a solid theoretical foundation, these definitions are accessible through SPARQL queries, contributing to a clear and coherent representation of knowledge.

[11] https://oasisbr.ibict.br/vufind/

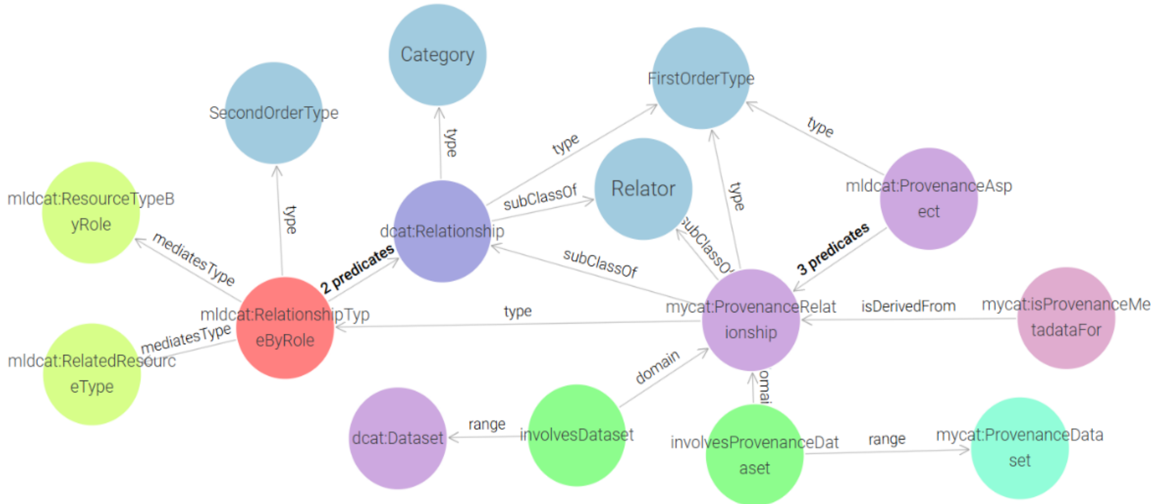
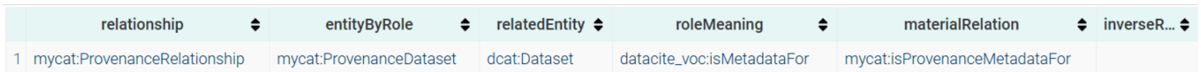


Figure 6: Relation and Relationship graph view using OWL.

SPARQL queries can be used to describe domain-specific relations, following the pattern for external descriptive relations. The formalism of the pattern enables the description of the entities involved, the relation, and its truthmaker. Figure 7 shows a SPARQL query that collects the elements involved in the pattern and the query results. This way, the schema, as a semantic model, provides relevant information for understanding domain-specific relations.

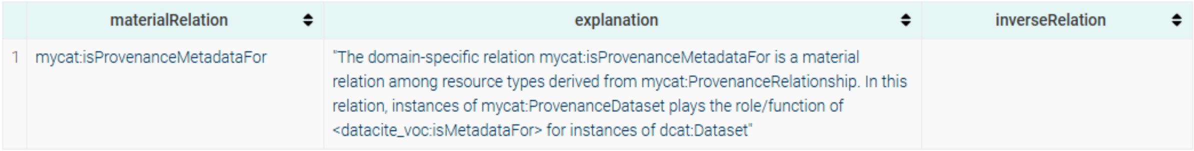
```
SELECT    DISTINCT    ?relationship    ?entityByRole    ?relatedEntity    ?roleMeaning
?materialRelation ?inverseRelation
  WHERE {
   ?relationship rdf:type/rdfs:subClassOf* mldcat:RelationshipTypeByRole .
   ?resourceRole mldcat:resourceRole ?relationship;
                  mldcat:hadResourceRoleValue ?roleMeaning .
   ?materialRelation gufo:isDerivedFrom ?relationship ;
                     rdfs:domain ?entityByRole;
                      rdfs:range ?relatedEntity .
   OPTIONAL
        {?inverseRelation owl:inverseOf ?materialRelation}
 } ORDER BY ?relationship
```



	relationship	entityByRole	relatedEntity	roleMeaning	materialRelation	inverseR...
1	mycat:ProvenanceRelationship	mycat:ProvenanceDataset	dcat:Dataset	datacite_voc:isMetadataFor	mycat:isProvenanceMetadataFor	

Figure 7: SPARQL Query to list domain-specific relations and relationships.

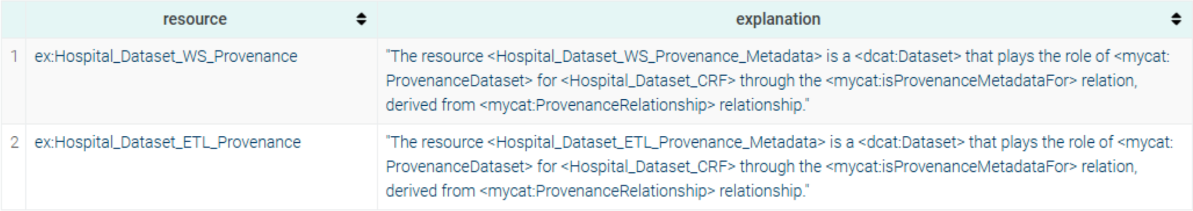
Recent research has experimented with the use of ontological patterns to explain the elements of ontology-driven conceptual models [16, 32]. Similarly, it is possible to extend the use of patterns to provide a view of the entities in the catalog schema. Therefore, by combining the information from the results in Figure 7 and using the pattern as a foundation, it is possible to design a SPARQL query to explain the material relations in the model. Figure 8 illustrates this example. Access to this information is useful for publishers who can select, from the relationships offered by the catalog, the one that best suits their needs or even ask the administrator to create a new one.



	materialRelation	explanation	inverseRelation
1	mycat:isProvenanceMetadataFor	"The domain-specific relation mycat:isProvenanceMetadataFor is a material relation among resource types derived from mycat:ProvenanceRelationship. In this relation, instances of mycat:ProvenanceDataset plays the role/function of <datacite_voc:isMetadataFor> for instances of dcat:Dataset"	

Figure 8: Explaining material relations in a Catalog.

The explanations can be extended to published descriptors. Thus, descriptors that assume a certain role because they are involved in domain-specific relationships may have an explanation associated with them, as shown in Figure 9. The figure shows an explanation for the workflow provenance dataset (Hospital_Dataset_WS_Provenance) and for the data transformation process provenance dataset (Hospital_Dataset_ETL_Provenance). Based on the pattern, it can be inferred that these resources are datasets that perform a specific role, defined as *mycat:ProvenanceDataset* for another dataset that contains patient clinical data according to the case record form (Hospital_Dataset_CRF). This function is represented by the relationship *mycat:isProvenanceMetadataFor*, derived from the relationship mycat:ProvenanceRelationship. Change logs stored as events could also be presented, demonstrating changes in provenance records over time.



	resource ⬍	explanation ⬍
1	ex:Hospital_Dataset_WS_Provenance	"The resource <Hospital_Dataset_WS_Provenance_Metadata> is a <dcat:Dataset> that plays the role of <mycat:ProvenanceDataset> for <Hospital_Dataset_CRF> through the <mycat:isProvenanceMetadataFor> relation, derived from <mycat:ProvenanceRelationship> relationship."
2	ex:Hospital_Dataset_ETL_Provenance	"The resource <Hospital_Dataset_ETL_Provenance_Metadata> is a <dcat:Dataset> that plays the role of <mycat:ProvenanceDataset> for <Hospital_Dataset_CRF> through the <mycat:isProvenanceMetadataFor> relation, derived from <mycat:ProvenanceRelationship> relationship."

Figure 9: Explaining dataset instances according to their role domain-specific relations.

The additional information provided by patterns offers insights into the semantic model of the catalog for agents. In particular, the truthmaker of relations gives a more precise understanding of the resources involved. Using SPARQL queries as semantic mechanisms, it is possible to validate new relations, ensuring the schema's consistency. These relations, formally and semantically expressed, can be reused by different resource publishers.

6. Conclusion and future work

This paper presented truthmaking and powertype patterns to enhance DCAT role-playing relations, thereby improving its semantics. As a result, instead of relying on embedded semantics at the instance level, a pattern was introduced to guide catalog administrators in creating dynamically domain-specific relations and their truthmakers at the schema level, contributing to the accurate representation of resources.

The introduction of multi-level types and an ontological foundation enhances the expressiveness and formalism of relations expressed using DCAT. Together, they provide means for defining, validating, explaining, and comparing domain-specific relationships. It is worth mentioning the treatment of highOTs as endurants, i.e., as entities with modal properties that can change qualitatively while remaining the same [33]. For instance, this enables the identification of resource types that play specific roles in relation to others. Software agents can handle the complexity introduced, making it transparent to researchers, catalog managers, and other users, as demonstrated with SPARQL queries.

A case in the area of clinical health data was used to demonstrate the use of an ontology that extends DCAT, adding highOTs that allow the definition of new relations and their relationships with their respective associated meanings. The publication of the data, along with the schema, in a triplestore such as GraphDB enabled us to illustrate the potential uses of ontological foundations and the adoption of ontological patterns in metadata schemas for catalogs. According to Auer [34], publishing the semantic model alongside the data represents an improvement in semantics, aiding in its contextualization. Additionally, the ontological foundation provided by UFO and its ontological patterns offers information about resource types and their relationships that is independent of any specific domain. This helps agents contextualize the domain and its data. The example demonstrates the potential for improving the quality of metadata schemas, which now have mechanisms to support communication with different agents. Using the semantic model, SPARQL queries can serve different human agents, from managers (e.g., catalog administrators) to data publishers and consumers.

This work, which focuses on improving relationships, is part of a research project aimed at enhancing the semantic expressiveness of DCAT by employing an ontological foundation and multi-level principles [35, 36]. This enhancement is essential for semantic interoperability, as shown by other studies in the field. In [37], for example, the authors incorporate key concepts from DCAT into the structure of the Elementary Multiperspective Materials Ontology (EMMO), a domain-specific ontology, to advance semantics for data sharing and use from the perspective of industry commons. In [38], the authors introduce the Data Catalog, Provenance, and Access Control (DCPAC) ontology. This ontology has DCAT and PROV-O at its core and combines other standardized ontologies and vocabularies to add a semantic layer to data lakes.

In the specific case of data catalogs that curate resource descriptors hosted on other platforms, the entities in the metadata schema classified as relators and events play a crucial role. They enable the management of changes that occur in descriptors over time. More than just a record as provided in *dcat:CatalogRecord*, which acts as a log for resources in the catalog, they establish a context for the changes, providing a provenance for the descriptors. In the future, this work will be expanded to address other types of relations according to the typology offered by UFO, further improving the expressiveness of the DCAT.

Acknowledgements

This work has been partially supported with research grants from RNP, CAPES (Process number 88887.613048/2021-00), and FINEP/DCT/FAPEB (nº 2904/20-01.20.0272.00).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Kropshofer, J. Schrott, W. Wöß, L. Ehrlinger. A Survey on the Functionalities of Data Catalog Tools. IEEE (2025). doi: 10.1109/ACCESS.2025.3568542.

[2] H. Sheridan, A. J. Dellureficio, M. A. Ratajeski, S. Mannheimer, T. R. Wheeler. Data Curation through Catalogs: A Repository-Independent Model for Data Discovery. JeSLIB (2021). doi: 10.7191/jeslib.2021.1203.

[3] S. C. Guptill. Metadata and data catalogues. Geographical information systems. (1999) .Vol. 2:677–92. URL: https://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/files/ch49.pdf

[4] D. Wells. Introduction to Data Catalogs. 2019. URL: https://www.ecintegrators.com.au/wp-content/uploads/2022/08/dave-wells-intro-to-data-catalogs-alation-2.pdf

[5] M. P. Satija, M. Bagchi, D. Martinez-Avila. Metadata management and application. LIBRARY HERALD. (2020). doi: 10.5958/0976-2469.2020.00030.2.

[6] R. Albertoni, D. Browning, S. Cox, A. N. Beltran, A. Perego, P. Winstanley. The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake. Data Intelligence. (2023). doi: 10.1162/dint_a_00241.

[7] O. Rouchon, E. Kraaikamp, E. Gonzalez, A. S. F. Kjeldgaard, N. P. Tenderup, J. Davidson, et al. D6.2 - Core metadata schema for legal interoperability. Zenodo (2024). doi: 10.5281/zenodo.11104269

[8] B. Van Nuffelen. Data Catalog Application Profile (DCAT-AP) - Version 3.0 [Internet]. SEMICEU (2024). URL: https://semiceu.github.io/DCAT-AP/releases/3.0.0/.

[9] A. Perego, B. van Nuffelen, GeoDCAT-AP - Version 2.0.0: A geospatial extension for the DCAT application profile for data portals in Europe, SEMIC Recommendation, European Commission, 2020. URL: https://semiceu.github.io/GeoDCAT-AP/releases/2.0.0/.

[10] L. O. B. S. Santos, K. Burger, R. Kaliyaperumal, M. D. Wilkinson. FAIR Data Point: A FAIR-Oriented Approach for Metadata Publication. Data Intelligence (2023); p. 163–183. doi: https://doi.org/10.1162/dint_a_00160.

[11] N. Guarino, T. P. Sales, G. Guizzardi. Reification and Truthmaking Patterns. Conceptual Modeling (2018). Cham: Springer International Publishing. pp. 151-65. doi: 10.1007/978-3-030-00847-5_13.

[12] F. Buschmann, K. Henney. Pattern-Oriented Software Architecture (2008). URL: http://www.dre.vanderbilt.edu/~schmidt/POSA-tutorial.pdf

[13] R. A. Falbo, G. Guizzardi, A. Gangemi, V. Presutti. Ontology patterns: clarifying concepts and terminology. In: Proceedings of the 4th International Conference on Ontology and Semantic Web Patterns (2013) - Volume 1188. Aachen, p. 14–26.

[14] G. Amaral, G. Guizzardi. On the Application of Ontological Patterns for Conceptual Modeling in Multidimensional Models. Advances in Databases and Information Systems (2019). Cham: Springer International Publishing. doi: 10.1007/978-3-030-28730-6_14.

[15] N. Q. Oliveira, V. Borges, M. L. M. Campos. Analysis of Observational Variables from an Ontological Patterns Perspective. Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO) (2022); URL: https://ceur-ws.org/Vol-3346/Paper5.pdf.

[16] E. Romanenko, D. Calvanese, G. Guizzardi. A Pattern-Based Approach for Explaining Ontology-Driven Conceptual Models. Intelligent Information Systems (2025). Springer Nature Switzerland, p. 137–44. (Lecture Notes in Business Information Processing; vol. 557). doi: 10.1007/978-3-031-94590-8_17

[17] N. Guarino, G. Guizzardi. "We Need to Discuss the Relationship": Revisiting Relationships as Modeling Constructs. Advanced Information Systems Engineering (2015). Springer International Publishing, p. 279–94. (Lecture Notes in Computer Science; vol. 9097). doi: 10.1007/978-3-319-19069-3_18.

[18] C. M. Fonseca, D. Porello, G. Guizzardi, J. P. A. Almeida, N. Guarino. Relations in Ontology-Driven Conceptual Modeling. Conceptual Modeling (2019). Springer International Publishing, p. 28–42. (Lecture Notes in Computer Science; vol. 11788). URL: http://link.springer.com/10.1007/978-3-030-33223-5_4.

[19] V. A. Carvalho, J. P. A. Almeida, G. Guizzardi. Using a Well-Founded Multi-level Theory to Support the Analysis and Representation of the Powertype Pattern in Conceptual Modeling. Advanced Information Systems Engineering (2016). p. 309–24. (Lecture Notes in Computer Science; vol. 9694). URL: https://link.springer.com/10.1007/978-3-319-39696-5_19.

[20] V. A. Carvalho, J. P. A. Almeida. Toward a well-founded theory for multi-level conceptual modeling. Softw Syst Model (2018) Feb;17(1):205–31. doi: 10.1007/s10270-016-0538-9.

[21] R. Albertoni, D. Browning, S. J. D. Cox, A. G. Beltran, A. Perego, P. Winstanley. Data Catalog Vocabulary (DCAT) - Version 3 [Internet]. W3C (2024). URL: https://www.w3.org/TR/vocab-dcat-3/.

[22] Unified Modeling Language, v2.5.1 [Internet]. Object Management Group (OMG); 2017. URL: https://www.omg.org/spec/UML/2.5.1/PDF.

[23] G. Guizzardi, G. Wagner. What's in a Relationship: An Ontological Analysis. Conceptual Modeling - ER (2008). p. 83–97. URL: http://link.springer.com/10.1007/978-3-540-87877-3_8.

[24] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. P. Sales. UFO: Unified Foundational Ontology. Borgo S, Galton A, Kutz O, editors. AO (2022) Mar 15;17(1):167–210. doi: 10.3233/AO-210256.

[25] A. Albuquerque, G. Guizzardi. An ontological foundation for conceptual modeling datatypes based on semantic reference spaces. IEEE (2013). pp. 1–12. URL: http://ieeexplore.ieee.org/document/6577693/

[26] G. Guizzardi, N. Guarino N. Explanation, semantics, and ontology. Data & Knowledge Engineering (2024). doi: 10.1016/j.datak.2024.102325.

[27] D. Costal, C. Gómez, G. Guizzardi. Formal Semantics and Ontological Analysis for Understanding Subsetting, Specialization and Redefinition of Associations in UML. Conceptual Modeling – ER (2011). p. 189–203. doi: 10.1007/978-3-642-24606-7_15

[28] G. Guizzardi. Ontological Patterns, Anti-Patterns and Pattern Languages for Next-Generation Conceptual Modeling. Conceptual Modeling (2014). p. 13–27. (Lecture Notes in Computer Science; vol. 8824). URL: http://link.springer.com/10.1007/978-3-319-12206-9_2

[29] R. A. Falbo. SABiO: Systematic approach for building ontologies. ONTO.COM/ODISE (2014). URL: https://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf.

[30] P. P. F. Barcelos, T. P. Sales, E. Romanenko, J. P. A. Almeida, G. Engelberg, D. Klein, G. Guizzardi. Inferring Ontological Categories of OWL Classes Using Foundational Rules. FOIS 2023. doi:10.3233/FAIA231122.

[31] G. L. Guidoni, J. P. A. Almeida, G. Guizzardi. Forward Engineering Relational Schemas and High-Level Data Access from Conceptual Models. Conceptual Modeling (2021). p. 133–48. (Lecture Notes in Computer Science; vol. 13011). URL: https://link.springer.com/10.1007/978-3-030-89022-3_12

[32] R. Confalonieri, G. Guizzardi. On the multiple roles of ontologies in explanations for neuro-symbolic AI. Mileo A, editor. Neurosymbolic Artificial Intelligence. (2025) Mar;1:NAI-240754. doi: 10.3233/NAI-240754.

[33] C. M. Fonseca, G. Guizzardi, J. P. A. Almeida, T. P. Sales, D. Porello. Incorporating types of types in ontology-driven conceptual modeling. Conceptual Modeling (2022). doi:10.1007/978-3-031-17995-2_2.

[34] S. Auer. Semantic Integration and Interoperability. Designing Data Spaces (2022). Cham: Springer International Publishing; p. 195–210. URL: https://link.springer.com/10.1007/978-3-030-93975-5_12

[35] V. Borges, E. M. Prata, M. L. M. Campos. Ontology-Driven Multi-Level Conceptual Modeling for Dataset and Distributions Descriptions. Frontiers in Artificial Intelligence and Applications (2023), doi: 10.3233/FAIA231132.

[36] V. Borges, N. Q. Oliveira, M. L. M. Campos. A Multi-level Ontology-based Approach for Descriptors of Catalogued Resources. Proceedings of the 15th Seminar on Ontology Research in Brazil (ONTOBRAS) and 6th Doctoral and Masters Consortium on Ontologies (WTDO), 2022; URL: https://ceur-ws.org/Vol-3346/Paper4.pdf.

[37] E. Ghedini, A. Hashibon, J. Friis. Semantic data exchange ontology. (2022). Zenodo. URL: https://doi.org/10.5281/zenodo.7784934.

[38] H. Dibowski, S. Schmid, Y. Svetashova, C. Henson, T. Tran. Using Semantic Technologies to Manage a Data Lake: Data Catalog, Provenance and Access Control. SSWS@ ISWC (2020). URL:http://www.ssws-ws.org/SSWS2020/SSWS2020_paper5.pdf

Modeling and Exploring Semantic View Metadata in Knowledge Graphs with VoSV*

Blind[1,2,*,†]

[1] *blind*

Abstract

A Knowledge Graph (KG) provides a semantic framework for integrating and managing diverse data organizations. This paper introduces *VoSV* (Vocabulary of Semantic View), a domain-independent vocabulary designed to annotate metadata in semantic views based on a structured Data Design Pattern (DDP). The DDP organizes data into a four-level hierarchical model, supporting scalable maintenance and context-aware exploration. At the core is the Metadata Graph, built using *VoSV*, which captures detailed, machine-readable metadata about structure, provenance, and quality. These semantic annotations enhance key governance functions such as data lineage, quality evaluation, and usability. The paper also presents an interactive tool for exploring metadata, allowing users to visually inspect metadata elements across multiple levels of the semantic view metadata graph. Together, the DDP, *VoSV*, and exploration tool promote transparency, accountability, and trust by offering a structured, semantic approach to documenting and managing data within KGs.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

A Knowledge Graph (KG) integrates different types of data sources to into well-grounded knowledge management, integration and intelligent analysis source. [1] cite that well formulated KG is really necessary when influencing large language models (LLMs). Though LLMs exhibit prowess in understanding and generating natural language, they mainly rely on probabilistic correlations in order to generate their outputs, which may lead to factual inaccuracies in the information in their output and also to semantic ambiguity and lack of specificity when the models are employed in specific domains. KGs, on the other hand, define the content explicitly in a structured and meaningful way, allowing it to be read by machines and such content is expressed in terms of the entities.

A KG ensures consistency, factual reliability, and contextual adequacy by provisioning access to verified information for concept disambiguation and maintaining logical coherence across the different semantic tasks that would, in turn, result in delivering applications that are more consistent and semantically stronger when applied to knowledge-intensive domains.

In this perspective, a semantic view of a KG provide a unified ontological framework emerging from the semantic integration of the data sources in a data lake [2]. This integration establishes a comprehensive and coherent organizational data environment, enabling seamless access and fostering streamlined decision-making processes.

The construction and maintenance of a semantic view in KGs present three major challenges: (i) extraction and transformation – integrating data from heterogeneous sources within a data lake into a unified representation, using a shared vocabulary defined by a semantic view ontology; (ii) semantic linking is the process of making connections between semantically equivalent entities from various sources so that they can align semantically; (iv) data fusion and quality improvement is the merging of

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

* You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

✉ blind (Blind)
🌐 blind (Blind)
iD blind (Blind)

many representations of the same real-world object into one logical representation to enhance data quality.

To address these challenges, [3] defines a DDP that supports all three (3) of the above mentioned elements of Data Design Patterns by organizing the data in KGs in a logical manner. The DDP will help to structure the data and metatada based on four distinct hierarchical data layers to support semantic data integration and provide easier ongoing maintenance of the semantic view in different contexts.

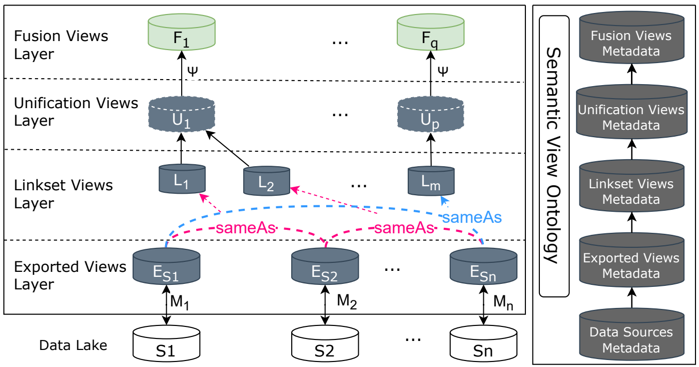
This paper introduces VoSV (Vocabulary of Semantic View), a domain-independent vocabulary, as a mechanism for creating comprehensive annotations for the metadata of semantic views created using the DDP. The metadata graph, which is the foundation of this architecture, is generated on the basis of VoSV. It also provides explicit and machine readable representations of the structure, provenance, and quality of the data in the semantic views. VoSV provides the basis for a number of important data governance functions, including the ability to track lineage, assess quality and improve the usability of data, through rich semantic annotations.

In addition, the paper also describes an interactive tool that will allow users to browse and retrieve metadata. By providing a structured and semantically grounded framework for documenting and governing digital resources, the DDP methodology, VoSV ontology, and metadata exploration tool will help to advance transparency and trust in digital ecosystems The DDP methodology, VoSV ontology, and Metadata Exploration Tool support visual inspection of various metadata types (e.g., provenance, quality) across multiple levels of granularity, creating a common language for documenting and governing digital resources.

The remainder of this paper is structured as follows. Section 2 introduces a four-level architecture for logically organizing data and metadata within a semantic view of a KG. Section 3 presents the core classes and relationships defined in the *VoSV* vocabulary. Section 4 describes the construction and exploration of the metadata graph for a semantic view modeled using *VoSV*. Section 5 discusses related works. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Data Design Pattern for Constructing a Semantic View

This section introduces a data design pattern, referred to as *DDP_SV*, specifically developed to logically organize both data and metadata within the semantic view [3]. At the core of the proposed framework, the semantic view is composed of two interconnected knowledge graphs: the **data graph** and the **metadata graph**. The data graph in Figure 1(a) represents the actual integrated data within the semantic view and works as the informational backbone of the knowledge graph. It is structured into a four-level hierarchical architecture, each level encapsulating a distinct stage of semantic integration:



(a) Data Graph **(b) Metadata Graph**

Figure 1: Data Design Pattern *DDP_SV* for KG's Semantic Views.

- **Exported Views Layer** – This foundational layer consists of RDF views generated by exporting data from raw sources in the data lake. Each exported view (EV) is obtained through an explicit mapping process that aligns the schema of a given data source (DS) with a shared vocabulary defined in the semantic view ontology (SVO). These mappings are specified by associating attributes from the DS with corresponding classes and properties in the SVO, typically through mapping languages such as RML/R2RML.
- **Linkset Views Layer** – This layer establishes semantic connections between equivalent entities across different exported views using identity relations such as *owl:sameAs*. The linkset views support semantic alignment and cross-referencing, forming the basis for further integration.
- **Unification Views Layer** – Built on top of the linkset views, this layer performs the integration of semantically equivalent entities into a single, canonical representation. The goal of unification is to ensure that all resources referring to the same real-world object under a common identity.
- **Fusion Views Layer** – The topmost layer addresses and resolves conflicts that may arise when the canonical representations produced by the unification views contain inconsistent or contradictory information. Building upon the aggregated representations produced by the unification views, this layer applies conflict resolution strategies to produce a consolidated and consistent view.

The metadata graph, depicted in Figure 1(b), serves as the repository for all metadata related to the semantic view. It plays a critical role in describing both the SVO and the views in all levels of the *DDP_SV*. Crucially, the metadata graph is semantically linked to the data graph, enabling integrated operations and contextual awareness. This tight integration supports a holistic understanding of the semantic view, empowering metadata-driven processes such as data discovery, quality assessment, lineage tracking, semantic governance, and view reuse.

3. Modeling Semantic View Metadata

This section presents the core classes of the *VoSV* (Vocabulary of Semantic View) and their interrelationships to represent the metadata of semantic views constructed using the *DDP_SV* design pattern.

Figure 2 provides an overview of the *VoSV*, highlighting its primary components and the semantic links between them. At the center of the model is the core class *vosv:SemanticView*, which conceptually encapsulates the main components of a semantic view. These components—and their respective roles within the metadata graph are described in the following subsections.

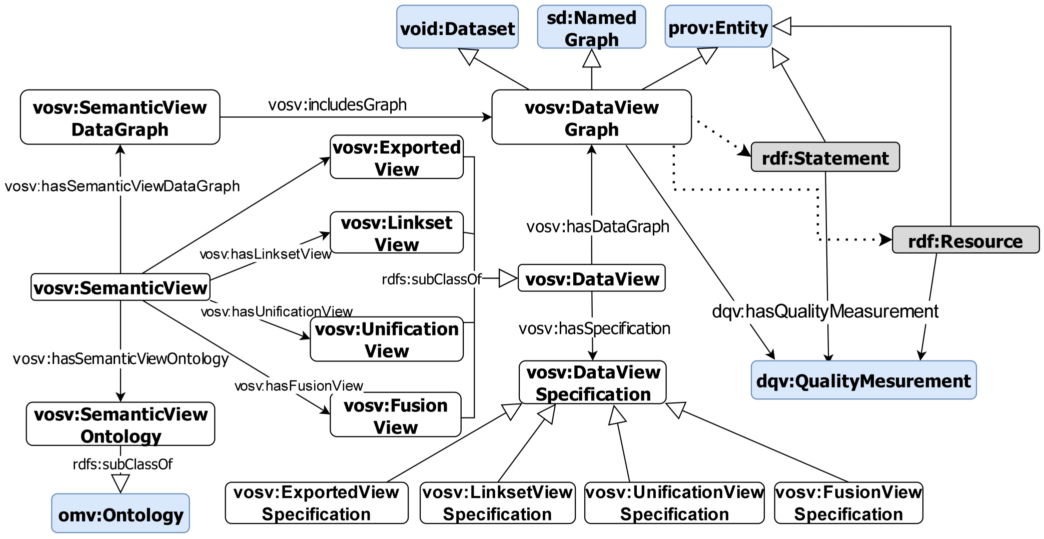


Figure 2: Overview of *VoSV* Ontology.

3.1. Reused Vocabularies

For modeling and construction of *VoSV*, standard W3C recommendation vocabularies were reused to allow the representation of metadata for elements of the knowledge graph construction process based on semantic integration. The Table 1 presents a list of reused vocabularies and their properties and concepts used in VoSV.

Vocabulary	Namespace	Properties / Concepts Used
Dublin Core	<http://purl.org/dc/terms/>	description, creator, created, format, issued, modified, language, publisher, contributor, subject, license
VOID	<http://rdfs.org/ns/void#>	Dataset, Linkset, vocabulary, dumpDataset, linkPredicate
PROV-O	<http://www.w3.org/ns/prov>	used, atTime, generatedAtTime, wasGeneratedBy, wasAttributedTo, Entity, Activity
PAV	<http://purl.org/pav/>	createdBy, providedBy, createdOn, lastUpdateOn, sourceLastAccessedOn, generatedAtTime
R2RML / RML	<http://www.w3.org/ns/r2rml#> <http://semweb.mmlab.be/ns/rml#>	TriplesMap, LogicalTable, LogicalSource, SubjectMap, PredicateObjectMap, source, template, sqlQuery, predicate
FOAF	<http://xmlns.com/foaf/spec/>	homepage
VANN	<http://purl.org/vocab/vann/>	preferredNamespacePrefix
OMV	<http://omv.ontoware.org/2005/05/ontology#>	Ontology, hasDomain

Table 1: Reused Vocabularies in VoSV.

3.2. Modeling the Semantic View Ontology

The semantic view ontology (SVO) serves as the foundational schema for constructing the semantic view of a KG. In the proposed framework, the SVO spans all layers of the *DDP_SV* and is built by integrating semantically equivalent elements across exported views.

In our incremental, pay-as-you-go approach, the SVO evolves progressively as new data sources are integrated. The initial version of the SVO mirrors the classes and properties of the first data source's ontology. When integrating a new DS, the SVO is extended to incorporate new classes and properties that accurately represent the semantics of DS. This integration involves two key processes:

- **Unification of Semantically Equivalent Properties** – Properties with similar semantic roles across different exported views are identified and unified under a common vocabulary. This standardization ensures a consistent vocabulary for describing attributes across all layers.
- **Establishment of Generalization Classes** – Semantically equivalent classes from various sources are grouped under shared superclasses, forming generalization classes. These provide an abstract semantic foundation for the unification and fusion views, supporting alignment at the conceptual level. This systematic approach ensures that the SVO remains scalable, maintainable, and semantically coherent, even as new data sources are integrated.

3.3. Modeling DataViews in the Semantic View

In the *VoSV*, all those individual views that compose a semantic view-namely exported views, linkset views, unification views, and fusion views-are modeled as instances of the class *vosv:DataView*. Each *vosv:DataView* is made up of two major components (see Figure 2)-a data view specification and a data view graph, which will be described in the subsections that follow.

3.3.1. Specification of Data Views with vosv:DataViewSpecification

Conceptually, the specification of a semantic view is defined as the union of the specifications of its data views. The class *vosv:DataViewSpecification* provides a declarative description of how a particular data view is constructed. It captures: (i) the input data sources utilized; (ii) the transformations applied to the data; and (iii) the structure of the resulting RDF triples.

To support specialization based on the four types of data views defined in the *DDP_SV* architecture, *vosv:DataViewSpecification* is further refined into the four subclasses described in the following (refer to Figure 2).

(i) Specification of Exported Views. The *vosv:ExportedViewSpecification* class defines the specification for exported views through three core properties:

- ***vosv:hasDataSource*** – specifies the source dataset from which the view extracts data;
- ***vosv:hasOntology*** – describes the classes and properties utilized in the exported view. This ontology fragment represents a subset of the SVO;
- ***vosv:hasMappings*** – defines the schema mappings and transformation logic connecting the source data to the RDF representation.

(ii) Specification of Linkset Views. The *vosv:LinksetViewSpecification* class captures the essential components of a linkset and is described by properties:

- ***vosv:hasSourceClass*** – identifies the source class (subject) from which links originate;
- ***vosv:hasTargetClass***– identifies the target class (object) to which links point;
- ***vosv:hasMatchFunction*** – describes the logic or rules used to establish semantic links (e.g., *owl:sameAs*) between source and target instances.

(iii) Specification of Unification Views. The subclass *vosv:UnificationViewSpecification* defines the specifications for unification views using two main properties:

- ***vosv:hasGeneralizationClass*** – specifies the generalization class (from the SVO) whose instances are to be unified;
- ***vosv:hasNormalizationFunction*** – defines the function used to canonicalize multiple IRIs representing the same entity into a single target IRI.

(iv) Specification of Fusion Views. The subclass to specification of fusion views is modeled by *vosv:FusionViewSpecification* , including:

- ***vosv:hasUnificationView*** – references the input Unification View that supplies the data;
- ***vosv:hasPropertyFusionAssertion*** – defines one or more property-level fusion assertions that resolve conflicting values for specific unified properties.

3.3.2. Representing the RDF Graph of a Data View with vosv:DataViewGraph

The class *vosv:DataViewGraph* represents the concrete RDF implementation of a data view within a semantic view. It encapsulates the actual RDF triples generated through either the materialization or virtual execution of the transformation logic defined in the corresponding view specification. Each *vosv:DataViewGraph* supports rich metadata annotations that describe its structure, provenance, and quality. To enable this, the class *vosv:DataViewGraph* is defined as subclass of the following foundational classes:

- ***void:Dataset*** – to support statistical and structural descriptions of the dataset, such as triple count, class usage, and property distribution;

- ***prov:Entity*** – to capture provenance metadata, including generation activities, responsible agents, and derivation history;
- ***sd:NamedGraph*** – to support SPARQL service descriptions, allowing the graph to be queried or accessed as part of a SPARQL endpoint.

In RDF, the triples contained within a named graph are effectively represented as quadruples, where the fourth element specifies the graph URI. This structure allows each triple in the semantic view to be explicitly associated with its corresponding *vosv:DataViewGraph*, enabling precise tracking of which view produced which data.

The class *vosv:SemanticViewDataGraph* represents the complete semantic view data graph of the semantic view, which is formed as the union of the *vosv:DataViewGraph* instances associated with all its constituent data views (e.g., exported, linkset, unification, and fusion views). The class *vosv:SemanticViewDataGraph* is defined as a subclass of *void:Dataset*, and each of its *vosv:includesGraph* properties directly references the IRI of a named graph, which corresponds to an instance of *vosv:DataViewGraph*.

3.4. Modelling Quality Metadata

Quality metadata plays a critical role in ensuring that a semantic view within a KG adheres to high standards of reliability, consistency, and trustworthiness [4]. These annotations capture key data quality dimensions—such as completeness, consistency, and coherence—and may also include constraints and validation rules designed to enforce data integrity.

The overall quality of a semantic view is assessed based on the data quality of its constituent data views, including exported, linkset, unification, and fusion views. To support a nuanced and scalable assessment process, data quality annotations are applied at four distinct levels of granularity:

- **Triple Level** – At the most granular level, individual RDF triples are evaluated to assess precision, correctness, or conflict indicators;
- **Instance Level** – The quality of a resource or instance is derived from the aggregated quality of its constituent triples;
- **Data View Level** – A data view's overall quality is evaluated by integrating the quality metrics of all instances within the data view;
- **Semantic View Level** – The quality of the complete semantic view is computed by aggregating the results of its component data views.

To ensure semantic interoperability and alignment with established standards, the *VoSV* adopts the class *dqv:QualityMeasurement* from the Data Quality Vocabulary (DQV)[5]. This class enables the representation of both quantitative and qualitative assessments of data quality. Each measurement is associated with a metric, which defines a standard procedure for evaluating a specific data quality dimension by observing concrete features of the data.

As illustrated in Figure 2, the classes *vosv:DataViewGraph*, *rdf:Resource* (representing individual entities), and *rdf:Statement* (leveraging RDF-Star to reference specific triples) are linked to instances of *dqv:QualityMeasurement* through the property *dqv:hasQualityMeasurement*. This design enables a structured and fine-grained approach to data quality evaluation across all levels of the semantic view, supporting detailed assessment at the graph, resource, and triple levels. By following this multi-level evaluation process, the *VoSV* provides a structured, extensible, and standards-compliant approach to data quality annotation.

3.5. Modelling Provenance Metadata

Provenance metadata captures information about the origin, derivation, and transformation of the data represented in the semantic view. It plays a critical role in ensuring transparency, accountability, and trust in integrated data by documenting where the data came from, how it was processed, and by whom.

Within the *VoSV*, provenance is consistently modeled across the different levels of the semantic view architecture defined by the DDP_SV. Specifically, provenance metadata is expressed at three levels: data view, resource and triple.

To ensure interoperability and adherence to established standards, the *VoSV* reuses the PROV-O[6] ontology to describe the provenance of data views graph, resources, and triples.

The class *vosv:DataViewSpecification* captures provenance information about the construction of each data view, such as the input data sources, the transformation or mapping strategies applied, and the declarative specifications used. This high-level provenance enables users to trace the data flow and decision points in the creation of each view.

The class *vosv:DataViewGraph*, which represents the implementation of the RDF graph resulting from the view, also includes provenance metadata—such as the agents responsible for generating the data (*prov:wasGeneratedBy*), and the activities that led to its creation (*prov:Activity*). This information supports auditability and enables verification of the trustworthiness of the produced graph.

At a finer granularity, provenance metadata can be computed and attached to individual resources and triples. Example, a resource of a fusion view can be annotated with details about which exported view resources it originated from and which fusion rule was applied. This level of detail is essential to enable fine-grained lineage tracking, auditable data integration, and knowledge construction.

4. Constructing and Exploring the Semantic View Metadata Graph

This section discusses the construction and exploration of the metadata graph of a semantic view modeled using the *VoSV* vocabulary. As a case study, we use the semantic view *SV_Music*, which integrates music-related data from two heterogeneous sources: *DBpedia* and *MusicBrainz*. These sources offer rich and complementary descriptions of musical artists.

4.1. Case Study: The Semantic View 'SV_Music'

Figure 3 illustrates a representative fragment of the semantic view *SV_Music*, which serves as the running example throughout the following sections. Figure 3(a) shows the metadata graph, annotated using *VoSV*, while Figure 3(b) presents the corresponding data graph.

This figure highlights the framework's capability to explicitly link metadata elements to the data they describe, enabling integrated exploration and contextual understanding.

In Figure 3, quality and provenance metadata are visually indicated by the icons 🎖 (quality) and 🧬 (provenance), respectively. These icons appear alongside specific resources and RDF triples within the semantic view, indicating the presence of associated metadata and facilitating rapid identification and exploration of key trustworthiness indicators.

4.2. Constructing the Semantic View Metadata Graph

This section illustrates the use of the *VoSV* to construct the metadata graph of a semantic view following the *DDP_SV*. The *DDP_SV* supports a "pay-as-you-go" strategy [7][8], enabling the incremental construction and maintenance of the semantic view as new data sources are integrated or updated over time. The approach consists of five main steps, which are detailed below.

4.2.1. Step 1 – Semantic View Ontology Modeling

As discussed in Section 3, the *DDP_SV* supports an incremental, pay-as-you-go approach to modeling the SVO, which involves two key activities: (i) the unification of semantically equivalent properties

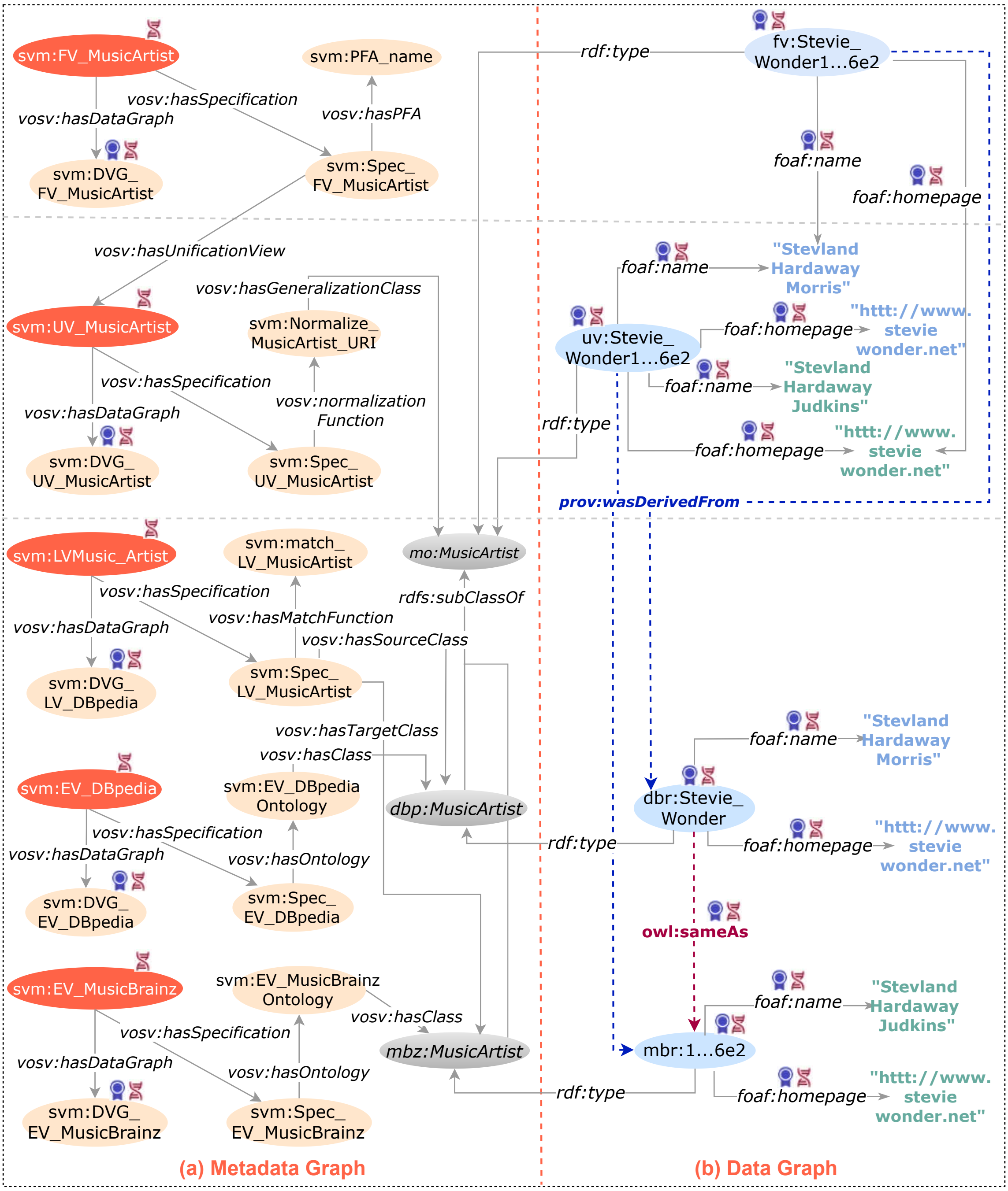


Figure 3: Fragment of *SV_Music* .

across different data sources and (ii) the establishment of generalization classes to represent common concepts shared across datasets. The detailed process of designing and evolving the SVO is beyond the scope of this paper.

The ontology of the semantic view *SV_Music*, named *SV_Music_OWL*, is constructed by uniting the vocabularies of data sources *DBpedia* and *MusicBrainz*. Figure 4 depicts, in UML, the main classes of the *SV_Music_OWL*. It reuses terms from three well-known vocabularies: *Dublin Core* (DC), *Friend of a Friend* (FOAF), and *Music Ontology* (MO).

SV_Music_OWL has the generalization classes *mo:Track*, *mo:Record*, and *mo:MusicArtist* to represent broader concepts encompassing more specific subclasses. For example, the generalization class *mo:MusicArtist* has the subclasses *dbp:MusicArtist*, and *mbz:MusicArtist*, which are specifically defined to annotate instances originating from the *DBpedia* and *MusicBrainz* data sources, respectively. This allows for tracking the provenance of the resources and visualizing the resource in different contexts.

The class *dbp:MusicArtist*, for example, is exclusively for the ontology of the *svm:EV_DBpedia*, meaning that instances of *dbp:MusicArtist* are derived from the *DBpedia* data source.

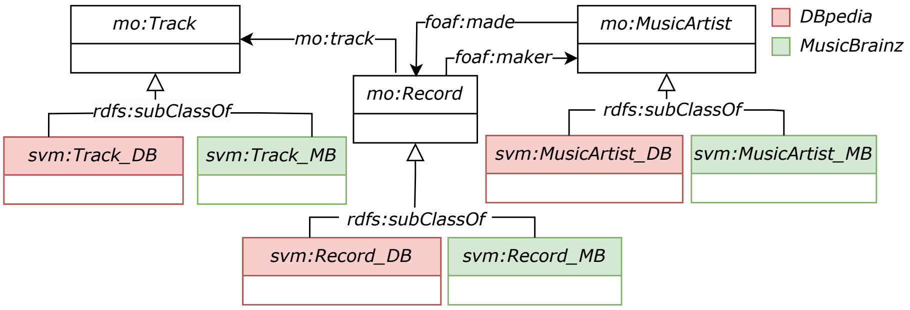


Figure 4: Main classes and relationships of SV_Music_OWL.

4.2.2. Step 2 – Exported View Construction

In this step, the objective is to construct an exported view for each data source by leveraging the existing semantic view ontology (SVO). The development of an exported view for a data source (DS) comprises two main tasks: (i) generation and validation of mappings from DS to SVO and (ii) implementation of the data graph, which can be virtual or materialized. The detailed process of constructing the export view is beyond the scope of this paper.

In the case study, the semantic view *SV_Music* includes two exported views: *svm:EV_DBpedia* and *svm:EV_MusicBrainz*. As illustrated in the metadata graph in Figure 3(a), each exported view is represented as an instance of *vosv:ExportedView*, encapsulating metadata that describes both its view specification and corresponding data graph. These data view graphs serve as access points to the RDF triples of each exported view and are implemented as distinct named graphs within the semantic view architecture.

4.2.3. Step 3 – Linkset View Construction

This phase focuses on creating linkset views to connect instances from an exported view to semantically equivalent instances in other exported views within a semantic view. The development of linkset views is carried out in three structured steps: (i) selection of source and target classes; (ii) definition of matching criteria; and (iii) creation of named graphs for linkset views.

Based on the class hierarchy defined in *SV_Music_OWL* in Figure 4, three linkset views are established to align semantically equivalent entities across sources: (i) *svm:LV_MusicArtist* – linking *dbp:MusicArtist* to *mbz:MusicArtist*; (ii) *svm:LV_Record* – linking *dbp:Record* to *mbz:Record*; and (iii) *svm:LV_Track* – linking *dbp:Track* to *mbz:Track*.

As shown in the metadata graph in Figure 3(a), the linkset view *svm:LV_MusicArtist* is modeled as an instance of *vosv:LinksetView*. It encapsulates metadata that defines both its view specification and the associated data graph, which together describe how identity links are established between musical artist entities from *DBpedia* and *MusicBrainz*. Figure 3(b) shows one example of sameAs triple (*dbr:Stevie_Wonder*, *owl:sameAs*, *mbr:1…6e2*) of *svm:LV_MusicArtist*.

4.2.4. Step 4 – Unification View Construction

In this study, a unification view must be defined for each generalization class present in the *SV_Music_OWL*. The construction of a unification view for a generalization class requires the definition of a “normalization function”. This function is responsible for remapping all IRIs that refer to semantically equivalent entities—declared as such via *owl:sameAs* links—into a single canonical IRI. The purpose is to unify multiple representations of the same real-world entity across different exported views.

In the case study, within the semantic view *SV_Music*, three unification views are constructed: *svm:UV_MusicArtist*, *svm:UV_Record*, and *svm:UV_Track*. Each unification view consolidates semantically equivalent instances from different data sources into a single, unified representation. For example, the unification view *svm:UV_MusicArtist* merges instances of *dbp:MusicArtist* from *svm:EV_DBpedia* and *mbz:MusicArtist* from *svm:EV_MusicBrainz*.

Figure 3(b) illustrates an instance of *svm:UV_MusicArtist*, where resources *dbr:Stevie_Wonder* (from *DBpedia*) and *mbr:1...6e2* (from *MusicBrainz*) are identified as semantically equivalent. These resources are normalized to a canonical IRI, *uv:Stevie_Wonder1...6e2*, which aggregates all classes, attributes, and relationships from both *dbr:Stevie_Wonder* and *mbr:1...6e2*.

The implementation of the data graph for a unification view is virtual. That is, the unification view is computed dynamically at query time rather than materialized and stored in advance. During query execution, all properties and relationships associated with semantically equivalent IRIs are consolidated and presented under the canonical IRI defined by the normalization function.

4.2.5. Step 5 – Fusion View Construction

This step focuses on implementing fusion views to resolve data conflicts arising from overlapping properties in the unification views. This involves defining fusion strategies that reconcile discrepancies and consolidate information accurately. Fusion views ensure that the most reliable and relevant data is presented in the semantic view, enhancing data quality and trustworthiness.

To construct a fusion view involves three main tasks: (i) identification of properties with conflicts information; (ii) creation of property fusion assertions; and (iii) implementation of the fusion view data graph.

As shown in the metadata graph in Figure 3(a), the fusion view *svm:FV_MusicArtist* is represented as an instance of *vosv:FusionView*. It encapsulates metadata specifying both its view specification and the associated data graph. To address conflicts for the property *foaf:name*, a property fusion assertion (PFA) named *svm:PFA_name* was defined as part of the fusion logic. Figure 3(b) illustrates the result of this fusion process for the canonical resource *fv:Stevie_Wonder1...6e2*, which is generated by consolidating information from the resources *dbr:Stevie_Wonder*(from *DBpedia*) and *mbr:1...6e2* (from *MusicBrainz*).

4.3. Exploring the Metadata of Semantic Views

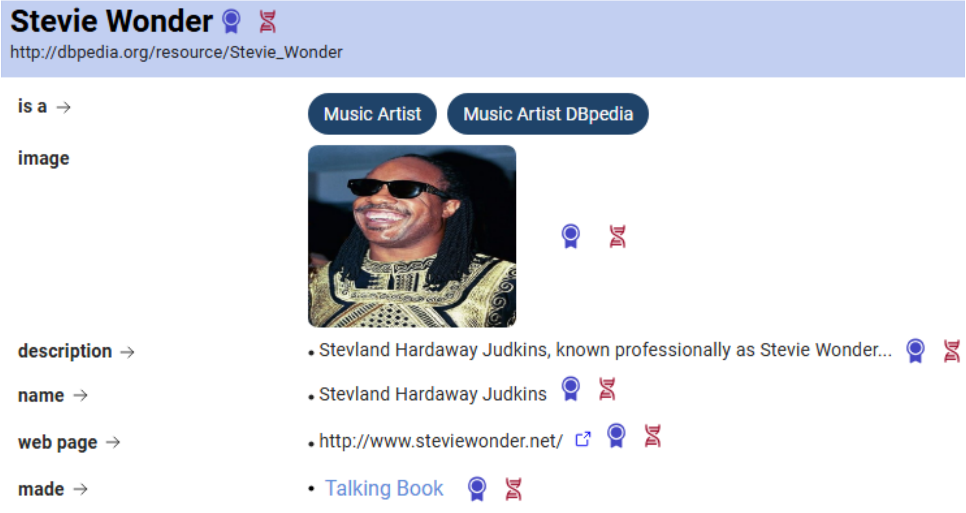
This section introduces a tool designed to support the interactive exploration and visualization of semantic view metadata. This tool is an enhanced version of the KG_Explorer platform [3], extended with new functionalities specifically developed for navigating and inspecting the metadata graph. These enhancements enable users to explore the structural, provenance, and quality metadata associated with each component of the semantic view.

In a semantic view constructed using the *DDP_SV*, as illustrated in Figure 3, resources and triples within the data graph are explicitly linked to their corresponding metadata in the metadata graph. This design enables integrated, contextual exploration of both data and metadata.

The tool allows users to inspect selected resources from the data graph in detail, including both their structural attributes and associated metadata. For example, Figure 5 displays the screen for visualizing the resource *dbr:Stevie_Wonder* within the data graph shown in Figure 3(a). The metadata displayed includes the resource's types—*dbp:MusicArtist* and the generalization class *mo:MusicArtist*—as well as the named graph to which the resource belongs: <http://blind-review/graph/ev_dbpedia>.

In the screen shown in Figure 5, icons displayed next to the resource's label and its properties provide direct access to the associated provenance and quality metadata, allowing users to quickly assess the reliability and origin of the data.

The tool retrieves quality and provenance metadata by querying the metadata graph and displays the information within the resource's exploration screen. Users can interactively select and explore



Stevie Wonder

http://dbpedia.org/resource/Stevie_Wonder

is a →	Music Artist Music Artist DBpedia
image	
description →	• Stevland Hardaway Judkins, known professionally as Stevie Wonder...
name →	• Stevland Hardaway Judkins
web page →	• http://www.steviewonder.net/
made →	• Talking Book

Figure 5: Exploration Screen for *dbr:Stevie_Wonder* resource.

the specific metadata elements they wish to examine. Quality metadata is retrieved by querying the precomputed quality measurements (*dqv:QualityMeasurement*) of data quality dimensions: consistency, completeness and timeliness. The computation of provenance metadata depends on the context of the resource–whether it belongs to an exported view, unification view, or fusion view as discussed below.

In the context of exported views, the provenance of a resource is retrieved by querying the metadata of the corresponding exported view, specifically its data view specification and the associated data view graph. The keys metadata retrieved include: (i) the data source that resource was derived from and (ii) the mapping rules that generated the resource. The provenance of individual triples is also derived from the metadata of the exported view specification, particularly by referencing the mapping rules defined for the triple's predicate.

In the context of unification views, the provenance of a unified resource comprises several key components: (i) the generalization class that defines the semantic type of the unified resource; (ii) the original resources from the exported views that were unified, referenced using *prov:wasDerivedFrom*; (iii) the normalization function applied to canonicalize the IRIs of the resources into a single, unified IRI within the unification view. This normalization process is modeled as a *prov:Activity*, and the unified resource is linked to it using *prov:wasGeneratedBy*.

In the context of fusion views, the provenance of a consolidated resource includes: (i) the generalization class; (ii) the original resources from the exported views that were merged (referenced using *prov:wasDerivedFrom*); and (iii) the property fusion assertions (PFAs) that resolve conflicting values at the property level. Each PFA is modeled as a *prov:Activity*.

The provenance of a triple in the fusion view context includes the original source triples from the exported views, which are referenced using *prov:wasDerivedFrom*, and the specific PFA responsible for resolving conflicting values for that property, which is referenced using *prov:wasGeneratedBy*.

5. Related Works

To the best of our knowledge, there is no existing vocabulary directly comparable to *VoSV* in its comprehensive treatment of semantic view metadata within KGs. Nevertheless, other vocabularies and models address specific aspects relevant to the representation and governance of semantic views, which was briefly review below.

The Vocabulary of Interlinked Datasets (VoID) [9], recommended by the W3C, is a widely adopted standard for describing RDF datasets and their interconnections. It supports metadata for dataset cataloging, discovery, and linkage. However, VoID primarily provides a static description of datasets and their links, and lacks the expressive capabilities required to model the structure, provenance, and transformation logic of semantic views within a KG. The Vocabulary for Attaching Essential Metadata (VAEM), proposed by [10], provides mechanisms to annotate metadata for artifacts such as ontologies, vocabularies, and services. It includes support for describing authorship, publication dates, licensing, and dependencies—thereby facilitating reuse and discovery. Despite its utility, VAEM does not address the modeling of semantic data views or their internal transformations, and therefore can not capture the semantics of view definitions or associated data graphs.

The Data Knowledge Vocabulary (DKV) proposes an ontology for documenting knowledge about data, including aspects such as provenance, generation processes, and data quality [11]. While DKV aligns with the goals of transparency and explainability in data pipelines, it treats datasets as monolithic entities. It does not model the internal decomposition into multiple views, nor the specific transformations, alignments, or fusion operations that underpin semantic integration in KGs. The work presented in [12] represents a significant effort in using knowledge graphs to document data integration processes. The author proposes a semantic data model capable of integrating ontologies, data sources, and mappings rules. While the conceptual approach aligns with the goals of *VoSV*, it is primarily limited to representing metadata related to exported views. As a result, it does not support the annotation of other types of semantic views—such as linkset views and fusion views—which are essential for capturing the full range of integration strategies in a knowledge graph.

In contrast, *VoSV* fills this gap by offering a vocabulary specifically tailored to annotate metadata across all levels of a semantic view, including structural and semantic specifications in semantic views within dynamic semantic integration environments.

6. Conclusions and Future Work

This paper addressed the problem of constructing and exploring the metadata graph of a semantic view within a knowledge graph. The semantic view is organized into a four-level hierarchical structure that supports modularity, scalability, and semantic traceability.

The paper presented *VoSV*, a domain-independent vocabulary specifically designed to annotate metadata across all levels of semantic views. The metadata graph modeled using *VoSV* offers a powerful foundation for efficient data management and utilization of the semantic view in KGs. *VoSV* enables structured and standards-compliant representation of key metadata—such as data quality and provenance—across multiple levels of granularity. It helps add detailed notes down to each resource and triple. This setup makes it easier to track where data comes from, explain how information is combined, and build knowledge that can be checked. With this aware design, the semantic views in the KG become more open, trusted, and manageable. They also stay verified, reusable, and well-kept even in large data systems.

To support practical usage, we have also shared a handy tool that helps visually check metadata letting users look into parts like provenance chains, quality notes, and structural metadata.

As future work, we intend to carry out empirical studies to evaluate the work, including also validating *VoSV* and KG_Explorer with experts in semantic knowledge graph. In addition, we envision *VoSV* being applicable across a wide range of scenarios. Among these, two key use cases reflect our current interests:

- Empowering semantic view construction with LLM agents – The *VoSV*-based metadata graph provides large language model (LLM) agents with a structured, reliable foundation that supports the automated synthesis of data integration pipelines [13]. By formalizing semantic view specifications and exposing them through machine-interpretable metadata, *VoSV* enables the development of agentic systems capable of dynamically constructing, validating, and managing semantic views in complex data integration environments.

- Establishing a source of trust in LLM-powered question answering – By integrating curated data sources, formal validation mechanisms, rich metadata documentation, and structured governance models, the metadata knowledge graph serves as a robust pillar of trust. In this context, *VoSV* transforms LLMs from generic text generators into guided and accountable agents—grounded in contextual semantics, capable of delivering context-aware, verifiable answers [14].

7. Declaration on Generative AI

No generative AI tools were used in the creation of the scientific content, data analysis, and presentation of results in this article. Generative AI was only employed, for minor language polishing and proofreading. All ideas, experiments, and conclusions are the sole responsibility of the authors.

References

[1] J. Sequeda, O. Lassila, Building enterprise knowledge graphs, in: Designing and Building Enterprise Knowledge Graphs, Springer, 2021, pp. 97–128.

[2] M. Galkin, S. Auer, M. E. Vidal, S. Scerri, Enterprise knowledge graphs: A semantic approach for knowledge management in the next generation of enterprise information systems, in: International Conference on Enterprise Information Systems, volume 2, 2017, pp. 88–98.

[3] V. Vidal, R. Freitas, N. Arruda, M. A. Casanova, C. Renso, A data design pattern for building and exploring semantic views of enterprise knowledge graphs, in: Anais do XXXIX Simpósio Brasileiro de Bancos de Dados, SBC, Porto Alegre, RS, Brasil, 2024, pp. 1–13. URL: https://sol.sbc.org.br/index.php/sbbd/article/view/30678.

[4] T. Hartmann, B. Zapilko, J. Wackerow, K. Eckert, Constraints to validate rdf data quality on common vocabularies in the social, behavioral, and economic sciences, arXiv preprint arXiv:1504.04479 (2015).

[5] R. Albertoni, A. I. E. Debattista, M. Dekkers, C. Guéret, D. Lee, N. Mihindukulasooriya, A. Z. Contributors, et al., Data on the web best practices: Data quality vocabulary (2016).

[6] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, Prov-o: The prov ontology, World Wide Web Consortium (2013).

[7] N. W. Paton, K. Christodoulou, A. A. Fernandes, B. Parsia, C. Hedeler, Pay-as-you-go data integration for linked data: opportunities, challenges and architectures, in: Proceedings of the 4th International Workshop on Semantic Web Information Management, 2012, pp. 1–8.

[8] J. F. Sequeda, W. J. Briggs, D. P. Miranker, W. P. Heideman, A pay-as-you-go methodology to design and build enterprise knowledge graphs from relational databases, in: The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Springer, Auckland, New Zealand, 2019, pp. 526–545. URL: https://link.springer.com/chapter/10.1007/978-3-030-30796-7_32.

[9] K. Alexander, R. Cyganiak, M. Hausenblas, J. Zhao, Describing linked datasets, in: Proceedings of the Linked Data on the Web Workshop (LDOW), 2009.

[10] R. Hodgson, I. Polikoff, Vocabulary for attaching essential metadata (vaem), https://lov.linkeddata.es/dataset/lov/vocabs/vaem, 2011.

[11] M. Böhmer, A. Dabrowski, E. Thordsen, Data knowledge vocabulary (dkv), https://lov.linkeddata.es/dataset/lov/vocabs/dk, 2017.

[12] S. Azizi, Documenting Data Integration Using Knowledge Graphs, Mestrado em ciência da computação, Gottfried Wilhelm Leibniz Universität, Hannover, 2023.

[13] D. Allemang, J. Sequeda, Increasing the llm accuracy for question answering: Ontologies to the rescue!, arXiv preprint arXiv:2405.11706 (2024). URL: https://arxiv.org/abs/2405.11706.

[14] J. Sequeda, D. Allemang, B. Jacob, Knowledge graphs as a source of trust for llm-powered enterprise question answering, Web Semantics 85 (2025) 100858. URL: https://www.sciencedirect.com/science/article/pii/S1570826824000441.

Applying principles of Ontology-Driven Conceptual Modeling to the interpretation process

Marcelo Jaccoud Amaral[1,*], Vânia Borges[1], João L. R. Moreira[2], and Maria Luiza M. Campos[1]

[1] *Programa de Pós-Graduação em Informática, Universidade Federal do Rio de Janeiro, Rio de janeiro, Brazil*
[2] *University of Twente, De Boelelaan 1105, 1081 HV Amsterdam, NL.*

Abstract

Effective interpretation of symbolic information is a critical challenge, especially as data grows in length and complexity, leading to an unpredictable or inadequately described surrounding context. Software development becomes a challenging task in the lack of a systematic description of the operating context, resulting in imprecise coding and ambiguous documentation. Interoperability standards, designed to align information meaning in heterogeneous environments, face similar issues. This work uses ontology-driven conceptual modeling to formalize the dynamic nature of interpretation as described by the Peircian Semiotics. It presents a generic model of the interpretation process that, by explicitly affirming its dynamic nature, directly contributes to the development of context-aware metadata standards. This formalization can significantly enhance the design of data models and standards, enabling the proper inclusion of contextual metadata, roles, and their dependencies, even within the limitations of current representation languages.

Keywords

Digital Object, Semiotics, Interpretation, Conceptual Model.

1. Introduction

In order to ensure information is reused in another system with the exact same meaning it was intended at the source, interoperability standards strive to register it with detailed documentation in clear language, and, more recently, with machine-readable semantic metadata. Standards for highly reusable data types, such as those directly supported by programming languages, have already coalesced into a few universally adopted ones, like Unicode [1] for text or IEEE 754 [2] for floating-point arithmetic. Most of these standards succeed in their task because they are designed to encode highly abstract types that are rarely used in their pure form, but rather as primitive types to build more complex and actually usable types.

For example, consider the words *cats* and *dogs*, which in most sentences, refer respectively to animals of the species *Felis catus* and *Canis familiaris*. However, when talking about animals in general, they may refer to the whole families *Felidae* or *Canidae*, which include wild cats like cougars and tigers, or coyotes and wolves. Consider now the meaning of the expression "It's raining cats and dogs". Now the same words convey a completely different meaning, that it is raining heavily. Note that this type of expression occurs in every language, and does not necessarily use the same words: in Portuguese, we would say "Está chovendo canivetes" (literally: it is raining pocket knives). Sometimes they are semantically linked: "Chove a cântaros" (lit. it rains in pitchers) correlates to "It's pouring rain" (from a vase). The context may influence the meaning of a word either subtly or radically. Consider also these two sentences:

[1] *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.*

[*] Corresponding author.

jaccoud@ufrj.br (M. J. Amaral); vjborges30@ufrj.br (V. Borges); j.luizrebelomoreira@utwente.nl (J. L. R. Moreira); mluiza@ppgi.ufrj.br (M. L. M. Campos)

0000-0002-1720-2245 (M. J. Amaral); 0000-0002-6717-1168 (V. Borges); 0000-0002-4547-7000 (J. L. R. Moreira); 0000-0002-7930-612X (M. L. M. Campos)

(a) The box could not pass through the window because it was too large.

(b) The box could not pass through the window because it was too small.

Notice that in (a) the pronoun *it* refers to the box, but in (b), the same word, in the same exact syntactic structure, refers to the window. This well-known linguistic problem is also linked to the context: the target of the pronoun is determined not only by the adjective ending the phrase, but also by the readers' knowledge about the domain of discourse.

This context dependence is not a distinct behavior of natural languages. It is present in computer languages (the plus sign may signify addition for numbers, concatenation for strings or character repetition in regular expressions), visual codes (red means danger in a road sign, spelling error in a syntax verifier, closing in the computer dialog window). Music, dance, and all forms of expression exhibit this same contextual behavior. As information grows in complexity, its dependence on context becomes more pronounced, often in ways difficult to formally describe. We believe the writing of documentation, and more broadly, interoperability standards, may benefit from a more detailed understanding of the interpretation process itself. As in most tasks where humans play a role, it is convoluted, complex, and sometimes impossible to automate. Already plagued by low-quality data, most diligent data managers have concerns about interpretations that may be incorrect, lossy, or imprecise.

In this paper, we aim to address these issues through ontology-driven conceptual modeling (ODCM), seeking a deeper understanding of the elements and mechanisms involved in the interpretation process. ODCM is the activity of capturing and formalizing how a community perceives a domain of interest using modeling primitives inherited from a foundational ontology [3]. It plays a fundamental role, helping us to understand, elaborate, negotiate and precisely represent subtle distinctions in our multiple conceptualizations of reality.

This work presents a generic conceptual model for the interpretation process. It aims to uncover the dynamic nature of interpretation, an aspect often disregarded in static relational schemas that assume data will always be used in the manner prescribed by data publishers. We believe that by using proper metadata to better describe the context in which data should be used, we end up with datasets that are more safely directed to software written for that scenario. To design such models, we need to understand the semiotic role of each property and its dependencies, considering the limits imposed by the representation, either human- or machine-related.

This paper is organized as follows: Section 2 presents background information on the terminology required for understanding this work. Section 3 addresses the interpretation process and presents an associated generic model. Section 4 explores this generic model by applying it to the interpretation of digital entities. Section 5 discusses requirements derived from the categorization of signs for metadata involved in the interpretation processes. Section 6 draws some conclusions and proposes directions for future work.

2. Background

The Semiotics of Charles Sanders Peirce (1839–1914) has been the object of many books and dissertations. The reader may refer to our short introduction in the discussion about the digital object concept [4]. A more detailed introduction may be found in [5], which used the Peircian concepts in his dissertation on Trinitology. We shall revise here the main topics regarding the process of interpretation, which Peirce called semiosis.

According to Peirce [6], the atomic piece for interpretation is the *sign*, which happens when something takes the place of something else in some context. Signs emerge continuously in our brains when cognitive signals from our senses match previously stored information regarding things they may represent. Later in his work, Peirce admitted that this process may also happen in the minds of other animals, or even in machines designed to reproduce the behavior of a mind. We interpret dark clouds as a rain forecast, a red light as a command to stop, a picture as the thing it depicts, and an utterance as corresponding words in a language. Signs usually come in chains: the smell of smoke may be interpreted as something being on fire, which in turn may be interpreted as danger, prompting us to run away or to seek means of extinguishing the fire. Signs do not emerge spontaneously: they require an intentional act by the interpreter agent. Things that happen mechanically, such as the production of electric signals by a microphone or a movement induced by a relay, occur without interpretation; they are simply physical reactions that happen in response to other phenomena. This is an important distinction to which we shall return later when regarding software.

Signs may vary largely in different scenarios. The well-known Saussurian linguistic sign is dyadic, composed of two parts: the signifier and the signified [7]. The Percian sign takes into account that the meaning may vary in different situations, and models the sign not as having 3 parts itself, but as something that emerges from the conjunction of a triad [6]:

- The *representamen* is the relevant aspect of something that depicts or describes something else, things that in common language we use to call symbols and graphic signs. Peirce specifically avoided the common terminology for this role in order to render it generic: any physical or mental manifestation may act as representation of something else.
- The *immediate object* is the relevant aspect of the thing being recognized. It is not an actual object in its plenitude, which is called the *dynamic object.*
- The interpretant embodies the knowledge of the situation, what links the representamen to the immediate object and the possible resulting effects. This is not the product of the sign, but what determines it. It is also not the interpreter, but depends on his competence.

Recalling the previous examples: dark clouds may be considered the representamen of a sign where imminent rain is the immediate object and the interpretant is the knowledge of the observer that the former usually precedes the latter; the redness of a traffic light is the representamen that signifies the danger of keeping on driving (the immediate object) which is determined by the driving knowledge (the interpretant) acquired by the driver and that results in braking the car. These three elements can be identified in every interpretation step.

It is important to note that the sign does not imply truth or correctness, since misinterpretations are also interpretations, and the interpretant is not limited to logical rules or formal knowledge. Some interpretations, and the actions they induce, are based on beliefs and convictions inherent in the agent. One common effect of a sign that has an assertion as its object is the storing of such information for future needs, changing the interpretant itself. Peirce called this recurrent sign-making process *semiosis* and believed this is how we acquire knowledge and pursue truth.

Another persistent element in Peirce's writings is the three universal categories, an abstract classification that he uses repeatedly in many situations. These trichotomies pervade his writings, and he purposely named the categories in a way that avoids confusing them with more concrete, quotidian categories. In a very simplified way, paraphrasing Peirce's words [8], these can be described as follows: *firstness* is what simply exists in itself, without referring to anything; *secondness* is what it is by force of something to which it is second; and *thirdness* is what it is through two other things that it mediates and brings into relation.

When we apply these categories to the sign model, we derive the three basic sign types: icon, index, and symbol. An *icon* is a sign in which the representamen displays firstness and relates to the object by a property of its own similitude, such as an image representing the thing it depicts. An *index* is a sign in which the representamen references or points to the object it represents, displaying secondness, such as a pointing finger, an arrow, an address, a compass needle that indicates North. A *symbol* is a sign in which representamen and object are bound by an arbitrary relation that mediates between them, displaying thirdness, such as words, ideograms, and all sorts of arbitrary rule encoding.

When the categories are reapplied to each type, we can derive other subtypes. We shall use these universal categories as a template to derive important metadata categories that expose these monadic, dyadic, and triadic valences.

3. Generic semiosis

As signs are very general in scope, a formal model of semiosis must address high-level entities in some upper-level ontology of choice. Because most of them define a description as a static relation, we had to resort to very abstract entities. We ended up choosing the Unified Foundational Ontology (UFO) [9], which is a grounded ontology that provides a foundation for domain analysis in conceptual modeling [10]. The UFO categories and relations enable the construction of a robust conceptual model that helps clarify concepts, facilitates meaning negotiation, and precisely defines the ontological semantics of represented notions [11]. UFO provides us with an adequate model for events and also an entity called relator, which is particularly useful in reifying complex relations like the ones involving a sign. UFO classes are expressed here with OntoUML, an ontology-driven modeling language that incorporates the distinctions underlying UFO into UML class diagrams [10]. It introduces various stereotypes (shown between «guillemets») that correspond to the concepts defined in UFO, as well as grammatical formal constraints that reflect UFO axiomatization [9], which allows us to better point out some analogies and patterns using a more concise notation.

To improve clarity, we have highlighted stereotypes related to UFO categories used in this paper for endurants and perdurants. Endurants are entities that exist in time with all their parts. They have essential and accidental properties and can undergo qualitative changes while maintaining their identity. Before proceeding, it is essential to discuss the concepts of sortal and non-sortal, as defined in UFO. Sortals are types that aggregate individuals with the same identity principle. In contrast, non-sortals aggregate individuals with different identity principles. In sortals, the identity is defined by a single «kind» it instantiates. The kind may have specializations that can be either rigid or anti-rigid. The former is stereotyped with «subkind» (e.g., hatchback car as a subkind of car). The latter classifies only contingently their instances and is stereotyped as «role» when grounded on relational properties (e.g., employee as a role of a person within the scope of an employment relationship). For non-sortal endurants, «roleMixin» defines contingent relational properties for individuals of multiple kinds (e.g., 'customer' for the kind person and organization). Intrinsic aspects of individuals, which depend on them or other individuals, are stereotyped as «quality» (e.g., car color) and «mode» (e.g., symptom). Extrinsic aspects are stereotyped with «relator» (e.g., marriage, enrollment). The «event» stereotype refers to perdurants, which are individuals that occur in time, accumulating temporal parts. They are existentially dependent on endurants that participate in them.

The main subject of a Peircian model of interpretation is the sign. We declared it as a perdurant, or, in OntoUML terms, gave it the «event» stereotype. It is important to stress that the sign, which denotes the meaning of something, is not a static object, but a temporal product of an intentional act happening in a particular situation. Even when a situation seems to repeat itself, the meaning may not be the same. Take, for example, the repetition of a mother's plea, "Go tidy your room!". Each utterance, even if using the same volume, results in a different interpretation. In this case, the same sign can never emerge again at a later point in time, because at least one of the components of the triad — the interpretant — is time-dependent. Even when the triad repeats itself, we consider the generated sign as another event, happening in a different time. This distinction between the triad and the sign is an important aspect of the Peircian model: the sign is not the triad, but emerges from it.

To clarify the roles played by the different components of the sign triad, we could just declare all related agents as participants. However, to better understand how each agent contributes to the sign emergence, we decided to reify the triad, a common modeling practice described in [12]. The result, however, is not a simple relator class. It not only brings together three very different participants, but their roles are characterized by complex dispositions embedded in each of them. Figure 1 depicts this general view and helps us unravel each component role separately. If no multiplicity is shown on an association end, it implies exactly 1. In the following text, expressions in **Capitalized Bold Font** refer to the ontology classes in the diagrams.

An important aspect of interpretation is that, being a process carried out by beings with a partial and imperfect view of the world, it works on this limited segmented information that is available to the interpreter. Peirce called these 'cuts' of the real things, and they apply to the three components, even if they are of distinct ontological natures. We have adjusted the terminology here to avoid confusion, since Peirce himself frequently used more than one term for the same concept.

The **Representamen** is "that character of a thing by virtue of which, for the production of a certain mental effect, it may stand in place of another thing" [13]. It summarizes the qualities of a larger thing, which, for practical purposes, we named a **Phenomenon**. A phenomenon is the object of the interpreter's perception and which has the potential for being interpreted as the occurrence or presence of something else, the signified thing. This is not restricted to physical phenomena, such as pictures or sounds, but also encompasses abstract or surreal ones, such as the shadows of monsters in a dark room.

The same cutting pattern happens with the signified component. In this respect, Peirce himself clearly differentiated between the actual full signified object, which he called **Dynamic Object**, and its qualities that are active in the triad, called **Immediate Object**. It is important to preserve this distinction because the actual dynamic object may be anything in the object of discourse, real or imaginary, and it is never known to the interpreter in its plenitude. We deal with and talk about things we only know in part, either by descriptions or direct inspection (their representamens).

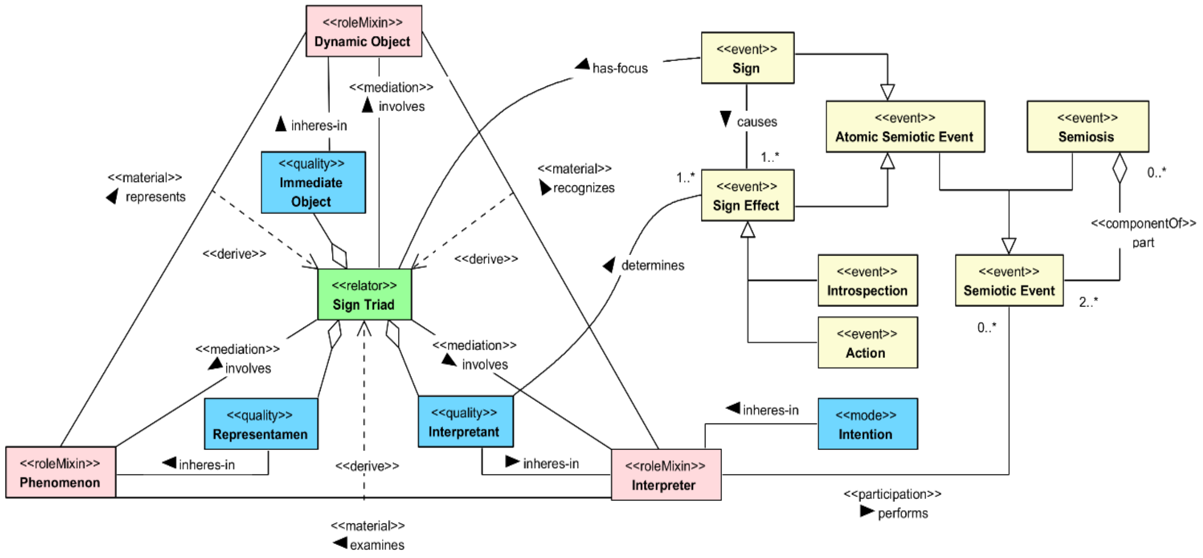


Figure 1: The generic sign model.

The third component is the actual difference from other semiotic models, since Peirce postulated that something can only be said to represent another in a particular context. A static signifier/signified pairing fails to capture innumerable situations where the context changes the meaning. The case of linguistics is exemplary: the same word or expression can mean many different things in different languages, different sentences, or different contexts. And may even signify two things at once. This is not necessarily an error; it may be intentional: a pun only achieves its goal (being funny or outrageous) if two meanings emerge simultaneously. A semiosis must account for all these strange kinds of interpretations. Peirce denoted this contextual part of the triad **Interpretant**, and although he also dissected this complex entity in many subtypes, we shall take a more pragmatic, task-oriented view of keeping the same pattern adopted for the other two components. We consider that, whichever qualities (modes or moments) the interpretant represents, they always inhere in the **Interpreter**, the real-world intentional agent who actually performs the task. These qualities may include:

- Knowledge of the relevant qualities in the representamen, which means acquaintance to important perceivable properties, such as physical modes (e.g. shapes or colors), symbolic codes (e.g. languages, their vocabulary, syntax or schemata), movements (e.g. in dance or sign languages), sounds (e.g. phones, tones, chords and intonation) etc.
- Knowledge of the relevant objects present in the domain, which may lead to the identification of suitable immediate objects.
- Time and spatial context: the meaning of words and gestures may vary radically from one decade to another, or from one place to another, from one user group to another.
- The source of the representamen. For example, the folded hands emoji (🙏) means please, thank you, praying, namaste, añjali mudrā or even a high-five, depending on where you are and with whom you are chatting.
- The principles used to match the representamen and the immediate object and determine the resulting effect of the emergent sign. Peirce frequently called these "habits", to avoid restricting them to logical rules. Interpretation may be based on stochastic functions, good or bad correlations, emotions, beliefs, etc.
- The history of the interpreter's dispositions. The agent's knowledge about a situation may change with time, resulting in different interpretations or actions.

This list may change drastically depending on the situation the interpreter is presented to. An agent who is experienced in translating languages may not have the proficiency to appreciate a complex piece of music or interpret a potentially dangerous situation while driving. The main intrinsic limitation with automating decision procedures using artificial intelligence is related to the colossal amount of information and modeling needed to support even simple scenarios.

Once the **Sign Triad** is formed, by matching the relevant parts in the three components, a **Sign** emerges, which in turn causes or evokes a **Sign Effect**, also determined by the interpretant. There are two basic types of effects: (i) the representamen is replaced based on the resulting immediate object (now in focus), creating a new matching step that will result in a new sign and a sign chain; we named this event **Introspection**, in analogy to the mental process of examining one's own thoughts; and (ii) the sign causes a physical action of some kind. If the resulting object is some assertion of a fact, the effect may be the storing of the information in memory for further use. If the result is a command of some kind, the agent executes it. This type of effect terminates the sign chain. This complex sequence of events constitutes the special process called **Semiosis**.

One should note that artificial symbolic representations, which act as representamens when they are perceived by another agent, are also the result of similar sign chains, in which the objects play different roles. For example, in reading, written characters are interpreted into spoken phrases, and in dictation the heard utterances are interpreted into written symbols. Note that they are not reverse processes; the signs and actions are not the same executed in reverse order, but, if executed correctly, they produce an inverse effect. As any foreign language student knows, it takes time to master both directions and perfect the different skills needed for each language.

This model is purposely abstract in order to be used as a template that must be specialized to reflect elements of the domain of choice. In the next session, we show how this may be applied to the interpretation of data.

4. Digital object semiosis and an application example

Applying Peirce's model to the world of digital computing may arise some doubts and misunderstandings. If the sign results from an intentional disposition from the interpreter, how would a machine with no volition perform a similar task? Souza [14], in her work on human-computer interaction, provides a clue to answering this question. She argues that our common perception of conversing with a user interface is, in fact, an illusion. Instead, the interface's responses are a direct consequence of how its human designers programmed it to logically react, mimicking anticipated human behavior. So, we are not actually talking to the machine, but talking to the designers via the machine, in the way they intended the conversation to unfold. We can say the machine acts as a delegate to the programmers, automatically responding to user inputs in the way the programmers would react. Also, a computer program does not spontaneously initiate; it is intentionally started by a human, either directly or indirectly. This also holds for very complex interfaces such as AI-driven assistants — there is a large infrastructure behind the interface, carefully constructed to behave in a human-like fashion, but the responses are all driven by the intentional acts of the many humans involved in its construction. So, if a piece of software is built to behave according to the will of one or many humans, we can say it inherits such intentional dispositions from the programmers. The software will always behave in the way it was programmed to behave and will do so because it is intentionally started and used. We can thus say that any interpretation that eventually happens by means of such code is intentional.

An important aspect of the digital environment is that it is very much uniform, since everything is represented by bit-encoded sequences. Because there is no natural source of data, every bit sequence present in the universe of a digital system comes either from some other system or from an analog-to-digital converter (ADC), a device that produces digital data from some analog sensor. For example, a microphone transforms sound waves into an analog signal, which is fed to an ADC circuit that regularly generates data encoded in some digital format like pulse code modulation (PCM). Also, everything that exists in the digital world needs a digital-analog converter (DAC) or similar circuitry that reacts to data stored in some memory and produces a physical reaction. For example, a sound DAC captures octets regularly from a serial interface and generates an analog electrical signal that may be later transformed into sound by a loudspeaker. So, our interpretation model deals exclusively with things that are derived from bit sequences. This is not a conceptual restriction, since in the generic model all components of the sign triad were also informational in nature, but a format restriction: in a digital system, everything *must* be represented by discrete bit sequences.

Porting our generic model into the digital space requires that each concrete class be replaced with some sort of subclass from a superclass we call **Bit Sequence.** To facilitate visualization, Figure 2 illustrates a model with the entities associated with this superclass, highlighting those involved in the process. Note that bit sequences and their subclasses are all of an informational nature, and these bits are always realized in some physical media, in this case, the computer memory. The derived model is illustrated in Figure 3.

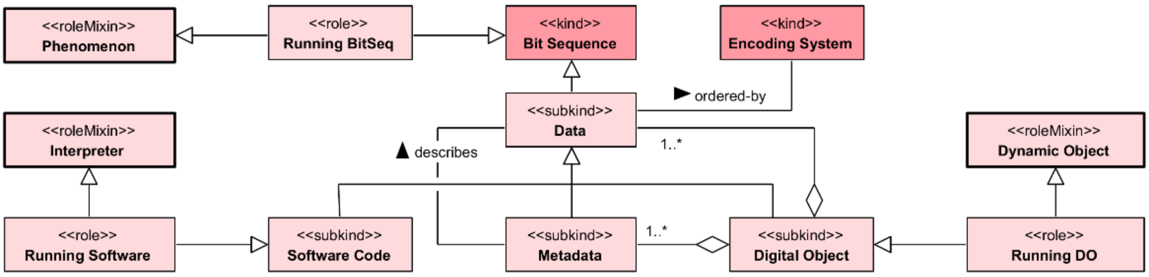


Figure 2: Bit sequence taxonomy and relations.

The definition of **Data** in this ontology is derived from [15], which borrows from DOLCE. It is a sequence of bits ordered according to some **Encoding System**. Although this may be used with all kinds of data, this example illustrates how it would work for a **Digital Object** (DO), which we define as in [4] as *"a finite bit sequence that is identified and has a set of assigned descriptors (also bit sequences) that support or aid in its interpretation"*. These descriptors were represented here by the common concept of **Metadata**, data that describes another piece of data. To simplify the example, we considered only DOs that are data, excluding those that may describe untyped or random bit sequences, which may appear in some scenarios.

Because the computer engine can only deal with bits, everything is a bit sequence that encodes some other entity. This includes descriptions of entities and all sorts of metainformation. Interpretation, thus, is the process of converting between representation spaces with different encodings. For example, in interpreting a program source code within a specific context (text encoding, programming language, syntax), the bit sequence for the string “0x34” is translated into a different bit sequence that represents the integer number *52*. Note that the computer does not actually know what *52* means, because it has no idea of what an integer number is. However, its circuitry knows how to do integer arithmetic on a limited range of integers by manipulating the bits with digital logic dates, and so it is able to do integer arithmetic without knowing what an integer is.

In fact, we have built software that can deal with thousands of other entity bit representations. But the computer can only correctly operate on data that is properly encoded, and representations must be constantly adapted in memory. Consider the example of software itself. It starts as ideas in a human mind, which are then encoded in text, the source code. Once the source code is input to a machine, it becomes a bit sequence, text encoded in some character encoding standard, such as UTF-8. The computer cannot run such a program encoded in text; it needs to interpret the text into an executable form, and it does so using another software called a compiler. The resulting bit sequence is now an object code, but still not ready to be executed. Another piece of software, part of the operating system, needs to copy such code into a specific memory address, adjust boundaries, and start a process to execute the code. The software is now a dynamic sequence of bits that continuously change (the quintessential Turing strip) until the program terminates. We use the same term — software — for all these bit sequences in different encodings, but they are distinct instances with different representations (source and object code) and even different ontological natures (a static plan, the program, and a mutable sequence manipulated by the running process).

Figure 3 shows the relevant parts of the interpretation process that, in different granularities, operate in such encoding transformations. The representamen is the relevant **BitSeq Aspect**s of the **Running BitSeq** that is being interpreted. When dealing with digital objects, the immediate object is the **DO Description**, which is the metadata associated with a particular **Running DO**. The interpretant is the context-aware **Algorithm** that the **Running Software** realizes**.** This triad is here a form of pattern match that results in the sign, the emergence of a **DO Occurrence**. The same algorithm determines which **DO Command** will result as an effect. This sequence of occurrences represents the **Software Process.**

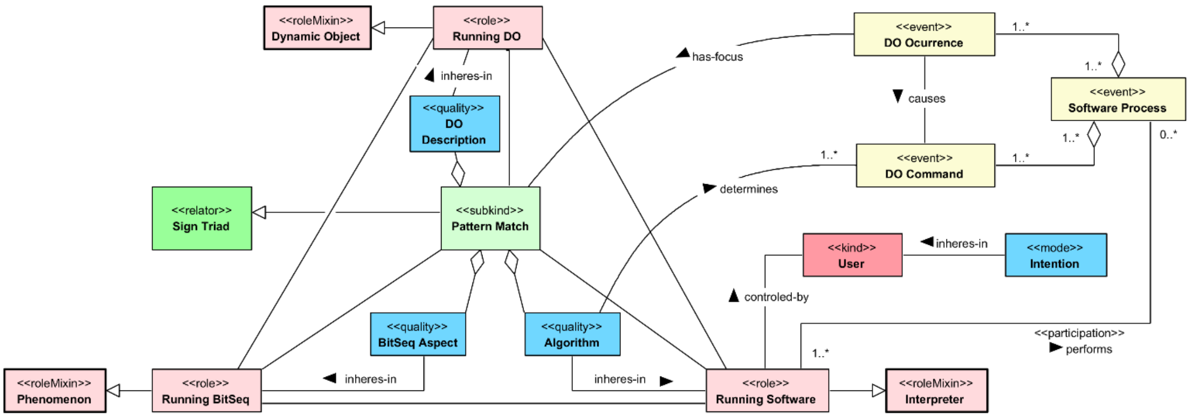


Figure 3: Digital sign model.

To help visualize this model applied to a real system, consider an XML parser, whose task is to interpret an XML-conforming text according to the semantics of an XML Schema. There are two possible approaches: the Document Object Model (DOM) parser [16] and the Simple API for XML (SAX) parser [17], as seen in Figure 4.

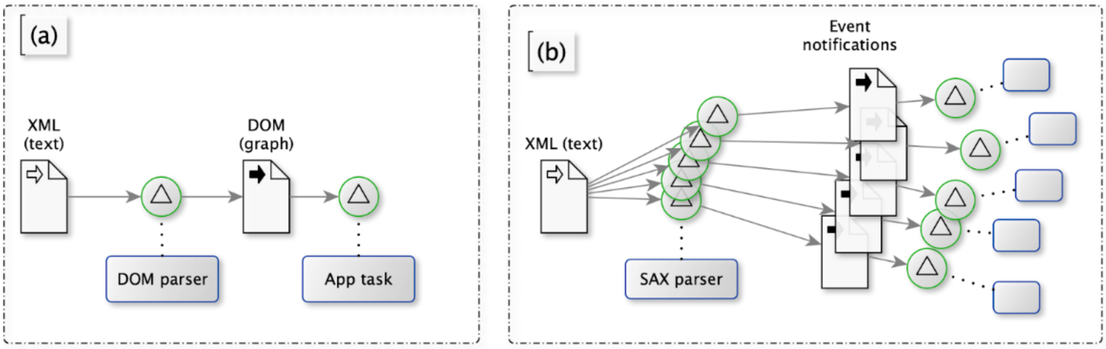


Figure 4: Example of DOM and SAX parsers.

The DOM parser (a) uses only its knowledge of syntax rules to determine and produce a new DOM graph, which is a new representation for the same information. Another routine, which knows nothing about XML, but needs to know the DOM encoding and the semantics inherent in the schema. The DOM sign focuses on the whole XML file at once, producing a single DOM graph, so the application has the whole context of the XML file to work with.

The SAX parser (b), which is used for very big files, translates every structural XML component (starting tag, attribute, ending tag, comment etc.) in discrete events, passing their representations as notifications to particular subroutines that will only be able to work with a smaller context of the XML tree. These are simpler signs, which result in less memory consumption and smaller routines, but are potentially restricted in interpreting the data because of the shrunken context.

In alignment to Peirce's view, parsing has already been considered as a knowledge acquisition method [18]. In a SAX parser, the schema and other previous information collected while traversing the XML tree may be cached to help resolve cross-references or improve the semantic resolution of terms. Every relevant information added to the interpretant should result in a better interpretation.

5. Metadata requirements

For each element of a data structure that requires interpretation, the interpretant must comprise knowledge of its ontological nature and the context for the operation. This information is provided statically by the software code or dynamically by metadata. Knowing how much metadata to provide is important to guarantee a proper interpretation. How much metadata is needed for each type of sign? Not coincidentally, there is a parallel with Peirce's sign typology and universal categories.

The universal categories may be applied to partition domains in segments that are characterized by the self (showing firstness), by something else (showing secondness), or by relating two other things (displaying thirdness). Dozens of such trichotomies appear in Peirce's writings, and many more have been developed by his disciples in many different domains. ***Table 1*** lists a few of a long list collected by [19] that are of interest to our analysis. The last three were used by Peirce to develop a more granular categorization of signs, which we have used to derive requirements for the metadata involved in the interpretation process.

Besides the already seen icon/index/symbol distinction, based on the way representamen and object are matched, Pierce considered how the sign is used: (i) a *qualisign* is based on intrinsic qualities that convey character, possibility, chance, disposition, not facts; (ii) a *sinsign* represents occurrence, reality, factuality, individuality, instantiation, and conveys haecceity and identity; and (iii) a *legisign* displays continuity, generality, relationship and changes in space-time. All symbols are legisigns.

Table1

Trichotomies based on the universal categories

Domain/Topic	**Firstness**	**Secondness**	**Thirdness**
Categories	Monads	Particulars	Universals
Modeling	Attributes	Individuals	Types
Predicates	Internal	External	Conceptual
Metaphysical concepts	Spontaneity	Dependence	Mediation
Relations	Attribute	External relation	Representation
Characterization	Uniqueness	Otherness	Composition
Inference	Emotional/Chaotic	Energetic/Stochastic	Logical/Formal
Signs	Representamen	Immediate Object	Interpretant
Sing-object	Icon	Index	Symbol
Sign use	Qualisign	Sinsign	Legisign
Interpretant mode	Rheme	Dicisign	Argument

The last trichotomy considers what drives the interpretant in the process. A *rheme* represents purpose and focuses on marks or the character of the representamen. A *dicisign* stands for fact, assertion, and, as such, focuses on the object. Only symbols and indexes behave this way, since icons are based on their own attributes. An *argument* expresses reason, rules, habits, and the influence of context. Examples are premises, conclusions, and a sequence of statements. Only symbols may act as arguments.

These sign trichotomies help us identify the type of knowledge required for the interpretant to perform its job, as well as the corresponding metadata. Table 2 outlines the metadata requirements in the case of digital objects.

Table 2

Metadata requirements by sign types

Sign subtype	Knowledge	Metadata
Icon	How bit sequences may be retrieved and compared. The “appearance” of a bit sequence is restricted by its length and the bit order.	None. The hardware is configured to handle bit sequences in a specific mechanical manner, and no interpretation is involved.
Index	How a bit sequence may refer to another one.	Standards for addressing bit sequences. Handled mainly by the operating system and network protocols. However, special standards may be developed for specific scenarios.
Symbol	How bit sequences encode other types.	Binary encoding standards. Formats and packing standards. Local defined types.
Qualisign	How are different possibilities encoded?	Knowledge of the possible arbitrary values an attribute may assume, which are provided by the Mode definitions.
Sinsign	How instances are qualified.	Standards for identifying resources, such as URIs, UUIDs, CRCs etc.

Sign subtype	Knowledge	Metadata
Legisign	How relations are encoded.	Standards for defining relations between objects, including the relations that classify them.
Rheme	How purpose is encoded.	Schemas and ontologies that characterize icons, indexes and symbols.
Dicisign	How facts are stated.	Languages used to describe entities, their types and associations. These affect only indexes and symbols.
Argument	How to reason.	Languages used to describe rules, laws, algorithms, behaviour, beliefs etc. in ways a CPU can replicate.

It is worth noting that the context dependence varies according to the complexity of the sign. On one side, qualisigns, which are based only on intrinsic properties, can only be iconic (nothing to index or mediate) and rhematic (fixed purpose), and thus require very little contextual information. Low-level routines in the software handle most of this low-level encoding, since the possible scenarios for bit encoding and memory manipulation are fixed and cannot be changed. On the other extreme, an argument needs many different standards to be efficiently interpreted, including the contextual information on which the algorithms apply. Some metadata annotation properties which are considered simple, like the Resource Description Framework (RDF) Schema comment[2] or the Simple Knowledge Organization System (SKOS) definition[3] are actually arguments that map from an artificial language (RDF or SKOS) into some natural language used to convey the intention of the author. Any system that needs to interpret such information will need a sophisticated interpretant armed with computational linguistic capabilities. This sort of metadata is clearly aimed at the human reader. Still, there is a lot of research using such information to help modelers in building or aligning ontologies [20, 21] or building natural language interfaces [22].

6. Conclusion and future work

The analysis proposed in this paper has revealed that various factors influence the interpretation of symbolic information encoded in digital systems. Although we automate this procedure with software and employ many methods to assure the quality of its intended execution, every exchange of data to or from another system implies a change of context that needs to be managed. It is not enough to guarantee that encodings and formats are compatible, but also that other contextual information is aligned, such as ontology versions, inference machines and rules. In the case of human interaction, localization parameters such as language and formatting are the same. This is especially important when data is exchanged, not in a neutral format, but one designed for human consumption.

The distinct aspects of knowledge required during interpretation depend on the various components of the sign triad, and this knowledge must be integrated into the running software for the system to provide the intended interpretation. This knowledge may already be embedded in the form encoding and format conformance, but it may also be loaded dynamically, in response to changes in the environment of accumulated knowledge. Proper validation of these new data is crucial to ensure the consistency of the running software. This is especially important with knowledge graphs, which cannot be fed conflicting information, and language models, which cannot receive their own output as feedback.

[2] https://www.w3.org/TR/rdf-schema/#ch_comment

[3] https://www.w3.org/TR/skos-reference/#notes

We hope this ontology helps the derivation of interpretation scenarios for other domains, but especially the case of proper metadata specification for digital objects, which lacks better validation procedures and for completeness and suitability. Incoherent metadata is an omen of incorrect data processing.

As future work, we intend to demonstrate how these principles can be applied to an RDF-based knowledge base and its associated services and show how we can derive Shapes Constraint Language[4] (SHACL) validations to pinpoint lags or inconsistencies within datasets, or problems resulting from the simultaneous use of multiple datasets with incompatible metadata.

Acknowledgements

This work has been partially supported with research grants from RNP, CAPES (Process number 88887.613048/2021-00), and FINEP/DCT/FAPEB (nº 2904/20-01.20.0272.00).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] The Unicode Consortium. The Unicode Standard, Version 16.0.0, (South San Francisco: The Unicode Consortium, 2024. ISBN 978-1-936213-34-4) https://www.unicode.org/versions/Unicode16.0.0/.

[2] IEEE Computer Society (2019-07-22). IEEE Standard for Floating-Point Arithmetic. IEEE STD 754-2019. IEEE. pp. 1–84. doi:10.1109/IEEESTD.2019.8766229. ISBN 978-1-5044-5924-2. IEEE Std 754-2019.

[3] T. P. Sales. Ontology Validation for Managers. (2014). 312 f. – Universidade Federal do Espírito Santo, Vitória, ES, 2014.

[4] M. J. Amaral, V. Borges, M. L. M. Campos. A Terminological and Semiotic Review of the Digital Object Concept, ER (2023). doi:10.1007/978-3-031-47262-6_5.

[5] A. Robinson, A. God and the world of signs: Trinity, evolution, and the metaphysical semiotics of CS Peirce (Vol. 2). Brill (2010). ISBN 978 90 04 18799 3.

[6] J. Hoopes (ed.). Peirce on Signs – Writings on Semiotic by Charles Sanders Peirce. The University of North Carolina Press (1991). https://muse.jhu.edu/book/41103.

[7] F. de Saussure. Curso de Linguística Geral (Cours de linguistique générale). Portuguese translation by A. Chelini, J.P. Paes and I. Blikstein. 28th. ed. Editora Cultrix (2012). ISBN 978-85-316-0102-6.

[8] C. S. Peirce, The Collected Papers of Charles Sanders Peirce, 8 Volumes, Cambridge, MA. Harvard University Press (1931) — p. 2.356.

[9] G. Guizzardi, C. M. Fonseca, J. P. A. Almeida, T. P. Sales, A.B. Benevides, D. Porello, Types and taxonomic structures in conceptual modeling: a novel ontological theory and engineering support. Data & Knowledge Engineering, (2021). doi: 10.1016/j.datak.2021.101891.

[10] G. Guizzardi, A. B. Benevides, C. M. Fonseca, D. Porello, J. P. A. Almeida, T. P. Sales. UFO: Unified foundational ontology. Applied ontology (2022). doi:10.3233/AO-210256.

[11] G. Guizzardi, H. A. Proper. On understanding the value of domain modeling. In: 15th International Workshop on Value Modelling and Business Ontologies, VMBO (2021).

[12] N. Guarino, T. P. Sales, G. Guizzardi. Reification and Truthmaking Patterns, ER (2018). doi:10.1007/978-3-030-00847-5_13.

[13] G. A. Benedict, What Are Representamens? Transactions of the Charles S. Peirce Society (1985). http://www.jstor.org/stable/40320088.

[14] C. S. de Souza, The semiotic engineering of human-computer interaction. MIT press (2005).

[4] https://www.w3.org/TR/shacl/

[15] D. Oberle. Semantic Management of Middleware. Springer (2006). ISBN 978-0-387-27630-4.

[16] T. Leithead. DOM Parsing and Serialization [Internet]. W3C (2016). URL: https://www.w3.org/TR/DOM-Parsing/.

[17] S. M. Foo, W. M. Lee. (2002). Simple API for XML (SAX). doi: 10.1007/978-1-4302-0829-7_7.

[18] A. Wallis, G. Nelson. Syntactic Parsing as a Knowledge Acquisition Problem. Proceedings of the 10th European Workshop on Knowledge Acquisition, Modeling and Management, EKAW'97, pgs. 285-300. Springer (1997).

[19] M. K. Bergman. A Knowledge Representation Practionary — Guidelines Based on Charles Sanders Peirce, p.143-7. Springer (2018). doi:10.1007/978-3-319-98092-8.

[20] B. Biébow, S. Szulman. TERMINAE: A Linguistics-Based Tool for the Building of a Domain Ontology Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management, EKAW'99, pgs. 49-66. Springer (1999).

[21] S. J. Ali, V. Naganathan, D. Bork. Establishing Traceability Between Natural Language Requirements and Software Artifacts by Combining RAG and LLMs. Proceedings of the 43rd International Conference on Conceptual Modeling, ER 2024, pg.295-314. Springer (2025).

[22] A. Oliveira, E. Nascimento, J. Pinheiro et al. Small, Medium, and Large Language Models for Text-to-SQL. Proceedings of the 43rd International Conference on Conceptual Modeling, ER 2024, pg. 276-294. Springer (2025).

Ontologia Pinakes: uma análise com Modelos de Linguagem de Grande Escala

Bruno Carlos da Cunha Costa[1], Greicy Kely C. dos Santos[1],
Ana Carolina Novaes de Mendonça[1], Gabriel Moraes de Oliveira[1],
Dayane Onaga Ferreira Machado[1,2] and Ana Carolina Simionato Arakaki[2,3]

[1] *Instituto Brasileiro de Informação em Ciência e Tecnologia, SAUS, Quadra 5, Lote 6, Bloco H , Brasília, DF, 70.070-912 , Brasil*
[2] *Universidade Federal de São Carlos, Rod. Washington Luiz, km. 235, São Carlos, SP, 13565-905 , Brasil*
[3] *Universidade de Brasilília, Asa Norte, Brasília, DF, 70910-900, Brasil*

Resumo

A construção de ontologias é essencial para a representação formal do conhecimento, mas permanece um processo complexo e dependente de especialistas. Os Modelos de Linguagem de Grande Escala (LLMs) surgem como alternativas promissoras para apoiar essa tarefa. Este estudoe exploratório avaliou o uso dos LLMs GPT-4o (OpenAI), Claude 4 Sonnet (Anthropic) e Gemini (Google) na geração automática de ontologias a partir do Documento de Especificação de Requisitos (ORSD) da Ontologia Pinakes, desenvolvida manualmente no Protégé para representar os serviços bibliográficos do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), com foco no Catálogo Coletivo Nacional de Publicações Seriadas (CCN). As ontologias geradas em RDF (*Resource Description Framework*) foram comparadas à Pinakes, tomada como referência. Os resultados indicam ganhos em prototipagem e cobertura conceitual, mas revelam limitações em coerência lógica, tipagem de dados e aderência a padrões bibliográficos.

Abstract

The Ontology construction is essential for the formal representation of knowledge but remains a complex process that relies heavily on experts. Large Language Models (LLMs) have emerged as promising alternatives to support this task. This exploratory study evaluated the use of GPT-4o (OpenAI), Claude 4 Sonnet (Anthropic), and Gemini (Google) for the automatic generation of ontologies based on the Ontology Requirements Specification Document (ORSD) of the Pinakes Ontology. The Pinakes was manually developed in Protégé to represent the bibliographic services of the Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), with a focus on the Catálogo Coletivo Nacional de Publicações Seriadas (CCN). The ontologies generated in RDF (Resource Description Framework) were compared against Pinakes, which served as the reference. The results indicate advantages in prototyping and conceptual coverage, but also reveal limitations in logical consistency, data typing, and compliance with bibliographic standards.

Keywords

Large Language Models, Ontology Engineering, Pinakes Ontology, Competency Questions,

1. Introdução

A engenharia de ontologias demanda tempo e conhecimento, desde a elicitação de requisitos até a modelagem formal e validação. Documentos de requisitos, como o *Ontology Requirements Specification Document* (ORSD), e conjuntos de Questões de Competência (QCs) são instrumentos consagrados para orientar esse processo [1, 2]. Com a popularização dos *Large Language Models* (LLMs), surge a possibilidade de empregá-los para apoiar a engenharia de ontologias, acelerando etapas iniciais [3],

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

✉ brunocosta@ibict.br (B. C. d. C. Costa); greicysantos@ibict.br (G. K. C. d. Santos); anamendonca@ibict.br (A. C. N. d. Mendonça); gabrieloliveira@ibict.br (G. M. d. Oliveira); dayanemachado@ibict.br (D. O. F. Machado); ana.arakaki@unb.br (A. C. S. Arakaki)

iD 0000-0001-9655-8698 (B. C. d. C. Costa); 0009-0008-3604-4298 (G. K. C. d. Santos); 0009-0004-0285-9932 (A. C. N. d. Mendonça); 0009-0005-4454-2340 (G. M. d. Oliveira); 0000-0001-5772-418X (D. O. F. Machado); 0000-0002-0140-9110 (A. C. S. Arakaki)

como a transformação de documentos de requisitos e QCs em rascunhos ontológicos, desde que a curadoria humana assegure qualidade e aderência semântica [4, 5].

Apesar do potencial, ainda há pouca evidência empírica sobre a eficácia dos LLMs em domínios especializados, como os serviços bibliográficos, nos quais predominam abordagens manuais. Nesse contexto, este estudo explora o uso de LLMs na geração de ontologias a partir de engenharia de *prompts* e do ORSD da Ontologia Pinakes, comparando as ontologias produzidas pelos modelos com a própria Pinakes, construída manualmente no *Protégé* [6], um software *open source* para desenvolvimento e validação de ontologias.

A Ontologia Pinakes é fruto de um projeto de pesquisa desenvolvido na Coordenação de Serviços Bibliográficos (COBIB) do Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), com o objetivo de promover a integração semântica dos dados e a interoperabilidade dos serviços tradicionais do Instituto. Sua construção baseou-se no modelo de domínio do Catálogo Coletivo Nacional de Publicação Seriada (CCN), alinhado a modelos conceituais de referência internacional, o que permite sua expansão para representar o conhecimento de outros serviços bibliográficos. Este estudo não teve como propósito validar a Pinakes, mas sim verificar a capacidade dos LLMs de interpretar requisitos formais e gerar estruturas ontológicas consistentes. A análise busca contribuir para o avanço da pesquisa em Ciência da Computação e Ciência da Informação, especialmente na geração automatizada de ontologias, com potencial de apoiar futuramente a modernização dos serviços bibliográficos do Ibict e o fortalecimento da infraestrutura semântica nacional.

2. Trabalhos Relacionados

A engenharia de ontologias é reconhecidamente complexa e demanda tempo, desde a elicitação de requisitos à formalização em OWL (*Web Ontology Language*). Nesse contexto, os LLMs têm sido investigados como alternativas para reduzir os esforços iniciais de modelagem.

Michael Funk, Jan Hosemann, Arne Jung e Carsten Lutz exploraram o uso do *GPT-3.5* para gerar hierarquias conceituais a partir de *seed concepts*, restringindo-se à estrutura taxonômica sem apoio de documentos de requisitos [3]. Anna Sofia Lippolis, Mohammad Javad Saeedizade, Robin Keskisärkkä, Aldo Gangemi, Eva Blomqvist e Andrea Giovanni Nuzzolese avançaram ao avaliar *GPT-4-1106*, *OpenAI o1-preview* e *LLaMA-3.1-405B-Instruct-16b* em um *benchmark* com dez ontologias, cem questões de competência e *user stories*, comparando as estratégias *Memoryless CQ-by-CQ* e *Ontogenia* [4]. Em trabalho posterior [7], os mesmos autores ampliaram os testes com LLMs de raciocínio, como o DeepSeek, incluindo formulação lógica além da geração de classes e propriedades. Mohammad Javad Saeedizade, Eva Blomqvist e Karl Hammar testaram técnicas de engenharia de *prompt* (*Zero-shot*, *Few-shot*, *Chain of Thought*, *Graph of Thought*, *Decomposed* e CQ-by-CQ) com GPT-3.5, GPT-4, Bard e LLaMA-2-70B, usando o *SPARQL Protocol and RDF Query Language* (*SPARQL*) como parte da avaliação da qualidade das ontologias geradas pelos modelos [5]. Em escala mais ampla, Jiayi Li, Daniel Garijo e María Poveda-Villalón realizaram uma revisão sistemática, destacando o uso recorrente da série *GPT*, *Claude*, *LLaMA* e *Mistral* em domínios diversos, como saúde, patrimônio cultural, finanças, gestão de emergências e organização do conhecimento acadêmico [8].

Apesar dos avanços, não foram encontrados estudos que empreguem LLMs na construção de ontologias voltadas à Biblioteconomia e aos serviços bibliográficos, com base em um Documento de Requisitos de Ontologia (ORSD) formalmente estruturado. Nesses domínios, predominam abordagens conceituais manuais, como o *IFLA Library Reference Model* [9] e o *PRESSoo* [10], aplicados inclusive na Ontologia Pinakes [6]. Projetos institucionais, como *BIBFRAME* [11] e a *Wikidata Bibliographic Task Force* [12], avançaram em direção ao *Linked Data*, mas com alinhamentos majoritariamente manuais. Catálogos internacionais, como o *British National Bibliography* [13] e o *BnF Data* [14], oferecem dados em RDF, porém sem geração automática via LLMs. Este estudo, portanto, contribui de forma inédita ao aplicar LLMs à geração de ontologias no domínio de serviços bibliográficos, tomando o CCN como caso.

3. Modelos Utilizados e Critérios de Escolha

A escolha dos modelos considerou critérios identificados na literatura e relevantes ao contexto institucional. Destacaram-se: (i) geração de saídas RDF/OWL válidas, demonstrada para GPT-3.5 e GPT-4 [5]; (ii) estabilidade e reprodutibilidade, frente a relatos de inconsistências em execuções [3, 5]; e (iii) diversidade de fornecedores e arquiteturas, fundamental para reduzir vieses e comparar abordagens distintas [4, 5, 7].

Também foram incluídos critérios pouco explorados, mas essenciais: (iv) custo, pela viabilidade institucional, e (v) suporte multilíngue, já que o domínio envolve recursos bibliográficos brasileiros alinhados a modelos internacionais. A relevância é reforçada por Li et al. [8], que evidenciaram a superioridade do GPT-4 em múltiplos domínios, por Lippolis et al. [4], que atestaram sua consistência estrutural, e por Luis Miguel Vieira da Silva, Aljosha Köcher, Felix Gehlhoff e Alexander Fay [15], que mostraram a capacidade do *Claude 3* e *Gemini Pro* de gerar axiomas *OWL* a partir de descrições em linguagem natural.

Para este estudo, foram selecionados o *GPT-4o* (*OpenAI*) e o *Claude 4 Sonnet* (*Anthropic*), pela precisão, raciocínio avançado e robustez conceitual; e o *Gemini 2.5 Flash* (*Google*) gratuito, por viabilidade de custo. Todos se enquadram nas versões já avaliadas na literatura, o que reforça sua relevância para fins comparativos. A intenção é analisar como cada um responde aos desafios da engenharia de ontologias, sem estabelecer hierarquia entre eles.

4. Metodologia

Este estudo exploratório analisa como diferentes modelos de inteligência artificial generativa interpretam e implementam um conjunto comum de requisitos ontológicos descritos no ORSD, tendo a Ontologia Pinakes como referência. Sem hipóteses rígidas, buscou-se identificar padrões e limitações ao longo de um processo dividido em quatro etapas.

A primeira etapa consistiu na definição do insumo de entrada para os LLMs, composto pelo ORSD da Ontologia Pinakes, cuja amostra está na Tabela 1, e por um prompt fornecendo contexto e orientações para geração da ontologia. Para os LLMs o ORSD incluiu: (i) especificações iniciais como objetivo, escopo e usabilidade; (ii) 59 Questões de Competência (QCs) com exemplos de respostas extraídas de registros reais do CCN; (iii) lista de termos derivados das QCs para enriquecer a terminologia; e (iv) quadro de alinhamento entre entidades dos modelos LRM, PRESSoo e CCN, produzido na fase conceitual da ontologia manual e incluído apenas como orientação estrutural para guiar os LLMs na geração de hierarquias compatíveis com padrões bibliográficos.

Na segunda etapa, de geração automática, o ORSD foi submetido ao *Claude Sonnet 4*, *Gemini 2.5 Flash* e *GPT-4o*, acompanhado do mesmo *prompt*, estruturado para gerar saídas em RDF/XML. Para fins descritivos, o conteúdo do prompt foi organizado em três seções: instruções iniciais, intruções complementares, com exemplos do domínio e orientações finais. Essa estrutura está resumida na Tabela 2 e aproxima-se da técnica *few-shot prompting* descrita por Lippolis et al. [7], diferenciando-se da abordagem tradicional ao utilizar um conjunto consolidado de QCs no ORSD e instruções finais para geração completa da ontologia. O *prompt* completo, o ORSD e todas as saídas RDF/XML estão disponíveis em repositório público[1].

Na terceira etapa, as ontologias geradas pelos LLMs foram carregadas no *Protégé* e comparadas à Pinakes. A análise comparativa considerou a cobertura conceitual, a organização hierárquica, o uso de metadados textuais e o esforço de edição. Para verificação preliminar de cobertura das QCs, estas foram convertidas em consultas *SPARQL*, um dos meios de validação de requisitos ontológicos, e executadas no *GraphDB*, um *triplestore* para dados em RDF. Os resultados dessa verificação foram tratados como indicadores, uma vez que o avaliador semiautomático, calibrado para a ontologia manual, pode ter descartado respostas corretas com variações de grafia ou estrutura. Paralelamente, na última estapa, conduziu-se uma análise qualitativa para identificar pontos fortes e limitações de cada LLM.

[1] https://github.com/cobib-ibict/OntologiaPinakescomLLMs

Tabela 1
Amostra de elementos incluídos no ORSD da Ontologia Pinakes [6], empregados como insumos semânticos para os LLMs

Componentes	**Exemplos (amostra)**
(i) Especificações iniciais	Objetivo: Desenvolver um modelo de dados em OWL flexível que represente o domínio dos serviços Bibliográficos do Ibict Escopo: O nível de granularidade está diretamente relacionado às questões de competência e aos termos identificados.
(ii) Questões de Competência (QCs)	QC1: "Qual é o título de uma determinada obra?" Resposta esperada: *Revista Ciência da Informação*.
(iii) Lista de termos	**Classes**: *PublicacaoSeriada*, *Biblioteca*; **Propriedades de Dados**: *tituloProprio*, *nome*; **Propriedades de Objeto**: *temISSN (hasISSN)*, *temEditora (hasPublisher)*;**Instâncias**: *Ibict*, *Ciência da Informação*.
(iv) Alinhamento LRM–PRESSoo–CCN	*LRM: Obra ↔ PRESSoo: Serial Work ↔ CCN: Publicação Seriada.*

Tabela 2
Estrutura resumida do *prompt* fornecido aos LLMs

Seção	**Elemento**	**Conteúdo resumido**
Instruções Iniciais	Objetivo e escopo	Gerar ontologia em *OWL* com base nos elementos do documento de especificação de requisitos, representando o domínio do CCN do Ibict, alinhada conceitualmente a LRM e PRESSoo; idioma pt-BR; saída válida para uso direto no *Protégé*
Instruções Complementares	Estrutura Taxonômica	Construir hierarquia com base no Quadro de alinhamento conceitual LRM–PRESSoo–CCN do ORSD.
	Exemplos estruturais	Exemplo de cadeia de especialização: *Biblioteca* $\subseteq$ *AgenteColetivo* $\subseteq$ *Agente* $\subseteq$ *Res*, onde $\subseteq$ indica relação de tipo (*subClassOf*).
	Diretrizes de modelagem	Criar todas as *classes* listadas no ORSD; criar *propriedades de dados e objetos* e *instâncias* listadas e extrair as não listadas a partir das QCs e de suas respostas; criar inversas quando aplicável; adicionar restrições de domínio e alcance e relações complexas quando inferíveis das QCs.
	Lista de termos	Exemplos de termos extraídos das QCs (e.g., Classes: *ISSN*, *Idioma*, propriedades de dados: *nome*, *tipologiaDocumental*, propriedades de objeto: *hasISSN*, *hasPublisher*, instâncias: *Ciência da Informação*, *Ibict*).
	Regras de nomenclatura	Classes em `CamelCase` (*PublicacaoSeriada*); propriedades em `camelCase` (*tituloProprio*); rótulos e comentários em pt-BR; evitar ambiguidade.
Orientações Finais	Instruções	Modelar apenas conceitos do ORSD, descrever anotações usando bases semânticas de recursos bibliográficos da web.
	Saída	RDF/XML: `OntologiaPinakes_[NomeLLM].owl`, pronto para cópia ou download.

5. Resultados e Discussões

A seguir, apresentam-se os principais resultados da comparação entre as ontologias geradas e a Pinakes construída manualmente, com destaque para estrutura, cobertura e desempenho dos LLMs.

5.1. Cobertura conceitual e estrutural

A Tabela 3 mostra que *Claude* e *Gemini* foram os modelos que mais se aproximaram quantitativamente da Pinakes manual, sobretudo no número de propriedades e instâncias. Já o *GPT-4o* apresentou cobertura limitada, mesmo após instruções reiteradas para maximizar o escopo.

Tabela 3
Comparativo entre os modelos gerados por LLMs e a Ontologia Pinakes manual

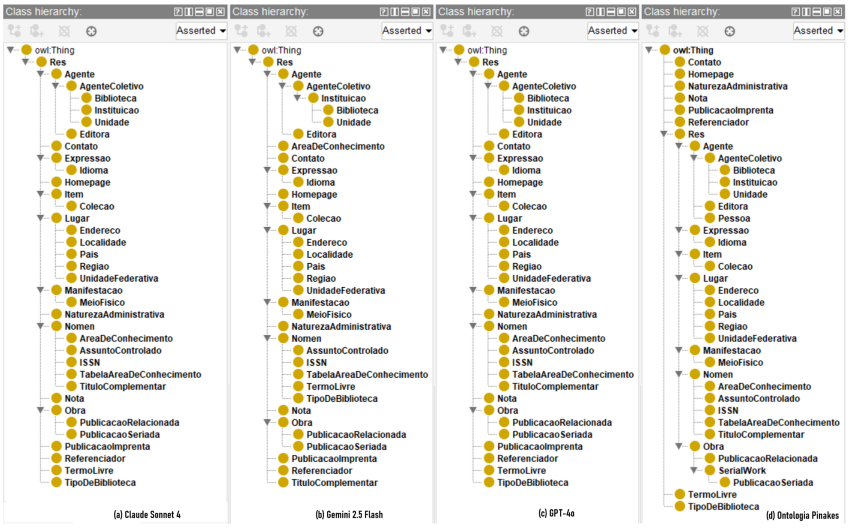
Modelo	Classes	Propriedades de Objeto	Propriedades de Dados	Instâncias
Claude Sonnet 4	36	52	45	63
Gemini 2.5 Flash	36	56	45	27
GPT-4o	36	40	34	24
Ontologia Pinakes	38	113	57	92

Apesar de gerar 56 propriedades de objeto, o *Gemini* deixou metade delas sem domínio e alcance declarados, o que comprometeu a precisão semântica, embora o raciocinador tenha inferido parte dessas restrições. O *Claude*, ao criar 63 instâncias, incorreu em excesso ao interpretar 18 valores literais como entidades independentes, inflando a contagem. Já o *GPT-4o* foi mais conservador, mas sacrificou cobertura: produziu menos de 25 instâncias e menos da metade das propriedades da ontologia manual.

Na análise preliminar de cobertura das QCs via consultas *SPARQL*, o *Gemini* apresentou a melhor correspondência ($\approx 58\%$), seguido do *Claude* ($\approx 51\%$), enquanto o *GPT-4o* ficou restrito ($\approx 27\%$). Embora esses percentuais ofereçam uma visão inicial do desempenho dos modelos, variações de grafia das respostas esperadas podem ter influenciado os resultados, especialmente por se tratar de consultas calibradas para a ontologia de referência.

5.2. Organização hierárquica e metadados

A comparação das hierarquias (Figura 1) mostrou diferenças relevantes. *Claude* representou múltiplos níveis, distinguindo *Obra*, *Manifestação* e *Item*, além de subdividir *Agente* e *Lugar*, aproximando-se do LRM. O *GPT-4o*, embora com menos elementos, preservou a estrutura central a partir da superclasse *Res* do LRM. Já o *Gemini* priorizou quantidade, mas com organização mais superficial. Nos metadados textuais, *Claude* e *Gemini* forneceram *labels* e *comments* semanticamente adequados, ainda que breves. O *GPT-4o*, em contraste, limitou-se a repetições pouco informativas, reduzindo a expressividade conceitual.



(a) Claude Sonnet 4

(b) Gemini 2.5 Flash

(c) GPT-4o

(d) Ontologia Pinakes

Figura 1: Comparação visual das hierarquias geradas pelos LLMs: (a) Claude, (b) Gemini e (c) GPT-4o, em relação ao modelo manual da (d) Ontologia Pinakes. Figura elaborada pelos autores, a partir da captura de tela do *software Protégé*.

Quanto às restrições, apenas o *Gemini* inferiu cardinalidades, embora algumas aplicadas incorretamente a literais. O *Claude* criou algumas inversas incompletas, corrigidas pelo raciocinador. Nenhum modelo gerou a classe *SerialWork* (PRESSoo), presente no ORSD, já a ausência da classe Pessoa (LRM) é compreensível, pois embora prevista para futura expansão da Pinakes, não foi incorporada ao ORSD por não estar alinhada às classes do CCN, o que limita sua inferência automática.

5.3. Aspectos práticos e limitações dos modelos

O *Claude* demonstrou boa capacidade de expansão conceitual, mas limitado pelo tamanho máximo das respostas, o que exigiu divisão do RDF em blocos. O *Gemini*, mesmo em versão gratuita, produziu uma ontologia relativamente robusta, mas sem exportação direta em RDF/XML. O *GPT-4o* foi o mais prático, permitindo importação imediata no *Protégé*, mas entregou menor cobertura conceitual.

De forma geral, todos os modelos respeitaram o escopo do ORSD e demonstraram ganhos expressivos na prototipagem, embora careçam de tipagem rigorosa e restrições adequadas, confirmando a necessidade de curadoria humana.

5.4. Observações sobre o esforço de modelagem

A construção manual da Ontologia Pinakes demandou dias de trabalho intensivo para modelar mais de 200 entidades, além da inserção manual de instâncias necessárias para validar as QCs, atividade trabalhosa e custosa de revisar. Em contraste, os LLMs geraram ontologias completas em poucas horas, dentro de um único dia de execução. Embora tenham exigido ajustes, especialmente em instâncias com atributos mal tipados e propriedades sem domínio ou alcance, o esforço global foi consideravelmente menor, evidenciando o potencial dos LLMs para acelerar fases iniciais da engenharia de ontologias.

6. Considerações Finais

A automação da construção de ontologias com LLMs mostrou-se promissora ao reduzir tempo e esforço nas etapas iniciais de modelagem. Este estudo evidenciou que LLMs geram classes, propriedades, instâncias e anotações válidas em RDF/XML a partir de um documento de requisitos (ORSD). Nenhum modelo, porém, alcançou integralmente a aderência hierárquica e conceitual nem a completude da Ontologia Pinakes. Seu uso ainda exige revisão e complementação especializada.

O trabalho contribui para o debate sobre a integração da Inteligência Artificial (IA) na engenharia de ontologias, destacando avanços e limitações. Como próximos passos, pretende-se explorar abordagens híbridas que combinem LLMs com extração automática de termos e relações, integradas a scripts em Python, além de aplicar outras técnicas de *prompt engineering* como *CQ-by-CQ* e *Chain of Thought* (CoT). Também será necessário avançar para aplicação em sistemas reais de gestão bibliográfica, avaliando o impacto na interoperabilidade e recuperação da informação, com aprimoramento do ORSD e do *prompt*, tornando-os mais claros, padronizados e alinhados às regras semânticas do domínio.

Declaração sobre IA Generativa

Os autores não empregaram ferramentas de IA Generativa; apenas aquelas inerentes ao objeto de análise do estudo.

Referências

[1] M. Grüninger, M. S. Fox, Methodology for the design and evaluation of ontologies, in: Workshop on basic ontological issues in knowledge sharing, Montreal, 1995. URL: https://www.researchgate.net/publication/2288533_Methodology_for_the_Design_and_Evaluation_of_Ontologies.

[2] M. C. Suárez-Figueroa, A. Gómez-Pérez, B. Villazón-Terrazas, How to write and use the ontology requirements specification document, in: Lecture notes in computer science, volume 5871, Springer, Berlin, Heidelberg, 2009, pp. 966–982. URL: https://link.springer.com/chapter/10.1007/978-3-642-05151-7_16.

[3] M. Funk, J. Hosemann, A. Jung, C. Lutz, Towards ontology construction with language models, in: Joint proc. of the 1st workshop on knowledge base construction from pre-trained language models (KBC-LM) and the 2nd challenge on language models for knowledge base construction (LM-KBC), co-located with ISWC 2023, volume 3577, CEUR-WS, Athens, Greece, 2023. URL: https://ceur-ws.org/Vol-3577.

[4] A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, Ontology generation using large language models, arXiv preprint arXiv:2503.05388 (2025). URL: https://arxiv.org/abs/2503.05388.

[5] M. J. Saeedizade, E. Blomqvist, K. Hammar, Navigating ontology development with large language models, in: European Semantic Web Conference (ESWC 2024), Springer, Cham, 2024, pp. 19–36. URL: https://link.springer.com/chapter/10.1007/978-3-031-60635-9_2.

[6] A. C. S. Arakaki, et al., Ontologia pinakes: formalização semântica para o catálogo coletivo nacional de publicações seriadas, in: Anais do VIII ISKO Brasil – 1º ISKO Brasil, 2025.

[7] A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, On the use of large language models to generate capability ontologies, in: Proceedings of the extended semantic web conference (ESWC 2025), 2025. URL: https://arxiv.org/abs/2503.08055.

[8] J. Li, D. Garijo, M. Poveda-Villalón, Large language models for ontology engineering: A systematic literature review, Semantic Web Journal (2025). URL: https://www.semantic-web-journal.net/system/files/swj3864.pdf.

[9] P. Riva, P. Le Boeuf, M. Žumer, IFLA library reference model: A conceptual model for bibliographic information, Ifla, Netherlands, 2017. URL: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf.

[10] PRESSOO Review Group, Definition of PRESSoo: A conceptual model for bibliographic information pertaining to serials and other continuing resources, version 1.3 ed., Ifla, Netherlands, 2017. URL: https://repository.ifla.org/items/4fcb0b59-1872-41ff-a9f9-1168bfb17d1a.

[11] K. Tharani, Linked data in libraries: A case study of harvesting and sharing bibliographic metadata with bibframe, Information Technology and Libraries 34 (2015) 5–19. doi:10.6017/ital.v34i1.5664.

[12] Wikidata, Wikidata:wikiproject bibliographies, 2025. URL: https://www.wikidata.org/wiki/Wikidata:WikiProject_Bibliographies.

[13] C. Deliot, Publishing the british national bibliography as linked open data, Catalogue & Index (2014) 13–18. URL: https://microblogging.infodocs.eu/wp-content/uploads/2014/10/publishing_bnb_as_lod.pdf, [Online].

[14] E. Grimaldi, The evolution of data.bnf.fr: Past, present and future of the bnf linked open data project, JLIS.it 15 (2024) 119–133. doi:10.36253/jlis.it-588.

[15] L. M. V. da Silva, A. Köcher, F. Gehlhoff, A. Fay, Assessing the capability of large language models for domain-specific ontology generation, in: Proceedings of the international conference on industrial ontologies (IO 2024), Springer, 2024. doi:10.1109/ETFA61755.2024.10710775.

Use of knowledge graphs applied to modeling mobile telecommunications network infrastructures in Brazil*

Daniele Nazaré Tavares[1,†], José Maria Parente de Oliveira[1,†]

[1]*Instituto Tecnológico de Aeronáutica (ITA), Praça Marechal Eduardo Gomes, 50. Vila das Acácias, 12228-900. São José dos Campos/SP - Brasil.*

Abstract

The information age and digital communication have transformed the way we collect, process, store, and display data. However, when dealing with georeferenced information, the challenge remains of representing spatial data in a manner consistent with its logical and semantic context. Conventional spatial tables describe attributes in isolation, failing to capture their meaning within the context in which they are embedded. This work proposes the use of knowledge graphs to integrate spatial and non-spatial databases, aiming to represent semantic relationships between geographic entities and telecommunications infrastructure across Brazil. Open data from Anatel and IBGE were used to structure the knowledge graph, which allows for the analysis of the distribution of radio base stations in each Brazilian municipality and their geoeconomic impact. The study demonstrates the feasibility of enriching analyses of mobile telecommunications infrastructure (2G, 3G, 4G, and 5G), supporting Anatel's decision-making in public policies aimed at ensuring interoperability.

Keywords

Ontology, Database, Knowledge Graphs, Knowledge Representation, Semantic Networks, Telecommunications, Mobile Networks, Geographic Information System (GIS)

1. Introduction

Mobile telecommunications network data is characterized by heterogeneous data types and complex relationships between the entities represented. However, the way in which data is currently available is limited in relation to these aspects. For this reason, a form of storage is needed that enables more efficient and meaningful organization of this data. Knowledge graphs are emerging as a promising alternative for organizing information on telecommunications infrastructure and services in Brazil.

In this paper, we present a proposal for a knowledge graph structure that represents the topology or architecture of mobile telecommunications networks in Brazil. The proposal allows for data scalability and more accurate analyses of the organization and behavior of radio base stations (ERBs) in different regions. The proposal also offers the following advantages:

- Knowledge graphs allow visualization of the network topology, understanding the relationship between its physical and logical components.
- With graph-based reasoning, it is possible to identify the root cause of failures or service degradation, streamlining the maintenance and restoration process of the telecommunication network.
- They allow the use of performance metrics related to network configurations to recommend antenna installation optimization strategies.

The geospatial database was obtained from open data from the IBGE (Brazilian Institute of Geography and Statistics), including names of states and municipalities, territorial geometries and other geospatial information about Brazil [1][2]. The data related to mobile telecommunications network infrastructure was extracted from the MOSAICO ou System Integrated Spectrum Management and Control System (Broadcasting Modules), made available by ANATEL (National Telecommunications Agency) [3].

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

dany.nt.14@gmail.com (D. N. Tavares); parente@ita.com.br (J. M. P. d. Oliveira)
https://danynazaretech.github.io/danynazare (D. N. Tavares)
0009-0001-2276-0145 (D. N. Tavares); 0000-0002-7803-1718 (J. M. P. d. Oliveira)

The rest of the document is organized as follows: **Section 2** presents related work in telecommunications that uses knowledge graphs and ontologies; **Section 3** describes the proposal for a knowledge graph for mobile telecommunications networks in Brazil; **Section 4** presents examples of queries made to the graph and discusses its benefits; finally, **Section 5** presents conclusions and suggestions for future work.

2. State of the Art

knowledge graphs began to be used in 2012 by Google [4] for the purpose of organizing data in the search for information at the conceptual level as a result of the interconnection of various types of data associated by semantic relationships. Knowledge graphs, which are data structures in graphs that combine data, relationships, and metadata to create a comprehensive understanding of the contextual object being represented, enable knowledge extraction [5][6]. This formal model is composed of nodes, which represent entities or objects, while the edges indicate the inference relationships established between them.

A notable example of this approach can be seen in Google's search engine. When you search for the name of a city, it not only displays related hyperlinks, but also presents organized information about its location, cultural aspects, health infrastructure, and geographic data. This data comes from an ontologically structured knowledge graph, which links different sources of information based on formally established semantic relationships.

In addition to applications in search engines, knowledge graphs have become fundamental tools in areas such as natural language processing, recommendation systems, semantic relation extraction, knowledge engineering, and intelligent chatbots. Their structure facilitates data organization during the preprocessing and modeling stages, making them especially useful for feeding artificial intelligence algorithms with contextualized and semantically enriched data.

The paper **Ontology for IP Telephony Networks** [7] presents an ontology in the field of IP telephony networks, focusing on the development of programs for this modalidality of network. The proposal is based mainly on the analysis of the SIP and H.323 protocols and the dominant signaling protocols in VoIP telephony. The telephony network structure is modeled with classes representing different types of nodes in the network: end-user terminals and network infrastructure. The goal is to standardize the telephony network structure and client usage to facilitate the development of telephony applications with interoperable data.

The TOUCA Project proposes the ToCo (Telecommunication CANvas Ontology) ontology [8], developed to represent hybrid telecommunications networks that combine wired and wireless technologies, various types of devices, interfaces, links, users, services, and channel quality in a modular and reusable way, based on the Device-Interface-Link (DIL) design pattern. ToCo allows for a semantic and integrated description of the infrastructure and behavior of modern networks, supporting applications such as software-defined networks (SDN), performance monitoring, and interoperability between technologies.

Advances in the use of knowledge graphs have also been made in the direction of efficiently compressing the density of data generated by 6G mobile networks, called pervasive multilevel native AI (PML)[9]. In this work, wireless data graphs are used to "*organize and condense complex and disordered data, extracting a concise subset that represents the most effective and critical factors for network AI models that process large volumes of wireless data*", contributing to the economy of memory in the storage of the data collected.

In the works presented above, it is worth highlighting the relationship between telecommunications networks and knowledge graphs to understand the semantic behavior between the physical and logical elements of a network for the construction of new protocols, architectural elements, topologies, and data refinement. In continental countries like Brazil, Russia, and China, where deployment costs are a barrier to its growth, this view of the network helps to build mobile telecommunications networks efficiently.

3. Knowledge graph of mobile telecommunications networks

Structuring spatial data together with non-spatial data is an inherently complex task, reflecting the wide diversity of formats, such as satellite images, cartography, vector data, matrix data, demographic data, textual data, and numerical data. Although tables help with organization, they are limited in their ability to visualize the semantic relationships between cartographic representation, geographic attributes, and external data from multiple sources. For this reason, geographic information systems (GIS) arose from the need to capture, manipulate, analyze, and model geographic data and metadata to provide an integrated and geospatial context-oriented view [10].

An additional limitation of the conventional tabular model lies in the difficulty of performing complex data queries, thereby reducing the potential to explore semantic relationships between data elements.

To represent the structure of mobile telecommunications networks in Brazil using a knowledge graph, the first step was to define the types of data to be used. Cartographic data in Shapefile format, numerical data, and text extracted from IBGE sources and georeferenced telecommunications data from Anatel's MOSAICO system were therefore used. Geographic, cartographic, and geoeconomic data from IBGE on Brazilian territory were used to analyze the structure of radio base stations (ERBs) with ANATEL data, interpreting how MOSAICO system mobile services are impacted throughout Brazilian territory.

Based on the understanding of this data, the corresponding knowledge graph was generated, linking the georeferenced data with the geographic data from the two databases. It expresses the relationships between the data of each ERB in relation to the spectrums it controls, its influence in the regional zone, antenna coverage by region, frequencies available for use, types of mobile services, and how each region manages each mobile telecommunications service, all with the organization of the geographic entities described in the IBGE data. Thus, as a basis for constructing the graph, the defined conceptual schemes of ANATEL and IBGE data show how the ERBs are distributed and their impact on each region.

3.1. Data description of Anatel and IBGE

The geospatial data extracted from the IBGE platform are in Shapefile (.shp) format, which contains the geometric characteristics and territorial boundaries of each Brazilian municipality. From the cartographic database, 5,572 municipalities and 27 states were extracted, along with six attributes for each municipality that are implicit in the map. External geographic information on states and municipalities was obtained from the Cidades@ portal [1], which describes the socioeconomic and structural characteristics of each locality, totaling 22,558 additional data points.

The SPECTRUM MOSAICO platform from Anatel is intended to manage telecommunications services, keep track of broadcasting station records, and show whether each ERB's infrastructure conforms with Anatel regulations. The open data used on licensed telecommunications stations in Brazil is in tabular .csv format with 2,083,046 rows and 41 columns, which separates the characteristics of each station's antennas. This study considers mobile service stations (2G, 3G, 4G, and 5G), and through their geographic position (latitude and longitude), it is possible to infer the impact of each ERB in the region where it is installed and how they interact with each other.

Using data from IBGE and Anatel, it is possible to establish spatial relationships with mobile architecture and its relationship with stations, making it possible to infer which antennas are in a specific area, perform network coverage or saturated area analysis, plan the network, and detect anomalies. In addition, it is possible to integrate data from other sources to infer the operational, administrative, or regulatory characteristics of public services. The principle behind this vision is the importance of semantic enrichment of data from multiple domains [10], so that end users can understand the impact of the mobile network across Brazil and consume it abruptly. .

3.2. Conceptual data diagram

A radio base station (ERB) is a fixed installation that houses the equipment responsible for enabling communication between mobile devices and the telephone company, providing signal coverage in a geographical area.

Based on current legislation [11] [12] [13], each ERB is defined in the graph as an entity according to the technical and regulatory criteria established by the MOSAICO system data. In the graph, each ERB is configured as a component responsible for the operational management of Antennas, presenting a 1:N cardinality relationship. Each Antenna is composed of physical attributes related to its operation, using wireless communication technologies such as 2G, 3G, 4G, and 5G. In addition, each one has concrete physical infrastructure and uses transmitting equipment to emit signals within the coverage area. Antennas installed in critical locations must be identified in order to follow an installation and operation protocol that is different from the others. Figure 1 shows the conceptual and topological model of the ERB structure with the antennas and technologies that operate, the technical characteristics under their operation, associated with the equipment used in the transmission infrastructure.

So each antenna has a concrete physical infrastructure and uses transmitter equipment to broadcast signals within the coverage area. Antennas installed in critical locations must be identified in order to follow an installation and operation protocol that is different from the others. The Figure 1 presents the conceptual model of the ERB structure with the antennas and the technical characteristics of their operation associated with the equipment used in the transmission infrastructure.

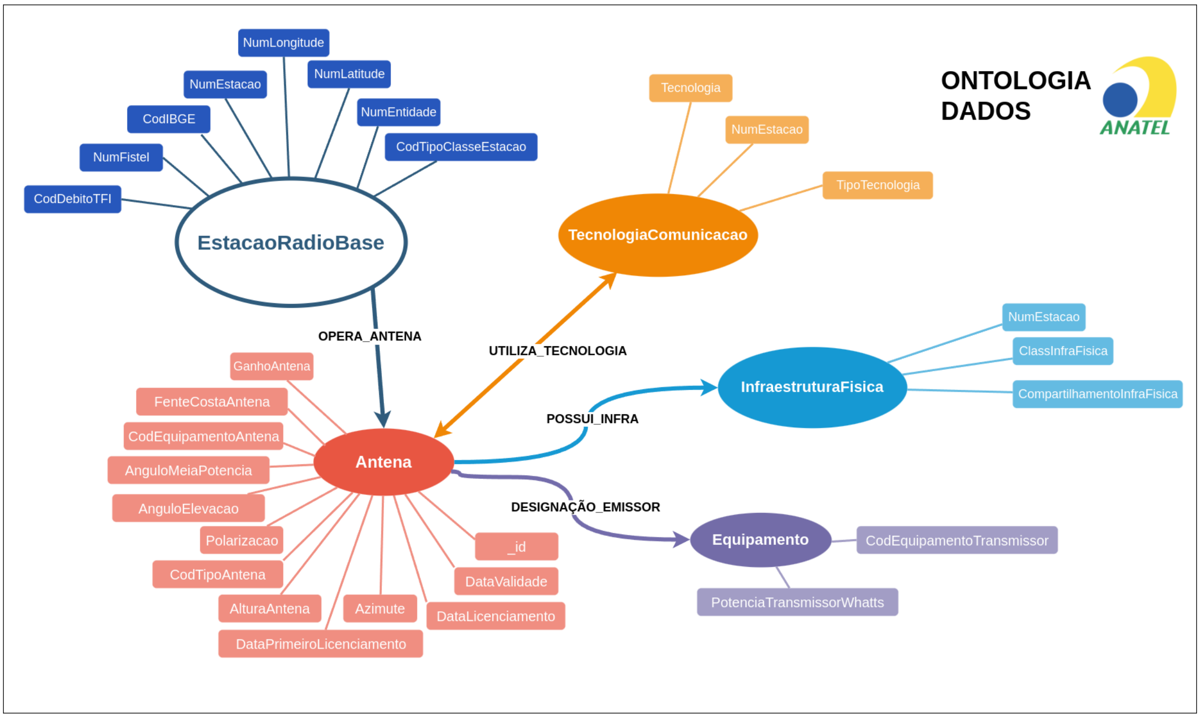


Figure 1: Conceptual model of the ERB structure.

To model the information related to licensed ERB concessions, latitude and longitude coordinates were used as a basis for establishing an indirect relationship between stations. This approach prevents data explosion if the ERBs were directly connected to other stations. With the support of thematic cartography extracted from the IBGE map, it is possible to indirectly associate each station with its structural characteristics and infer its area of influence in the geographical region where the ERB provides signal coverage.

Thus, to represent the distribution of ERBs in cities and organize them administratively in accordance with Anatel, only neighbouring cities in the same state were connected directly, and the state module was connected to the capital city, as illustrated in Figure 2. This connection between cities is useful for solving the travelling salesman problem when Anatel's regional superintendence needs to perform maintenance and inspections of its internal state mobile infrastructure. The states were connected directly to their neighbours to formalize the connections between states, according to IBGE data.

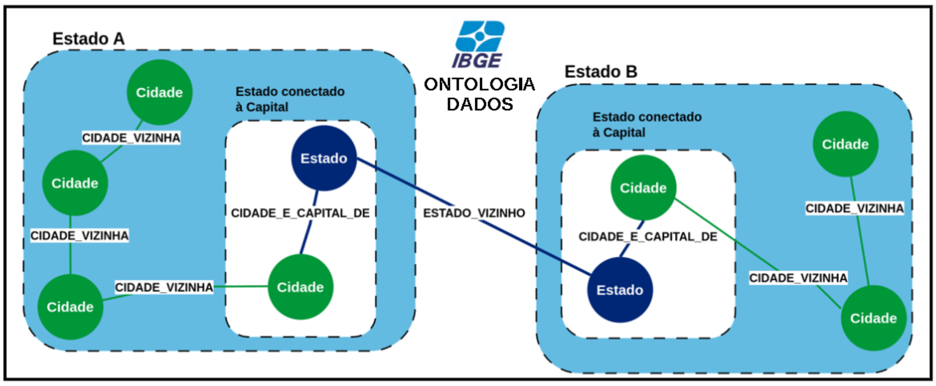
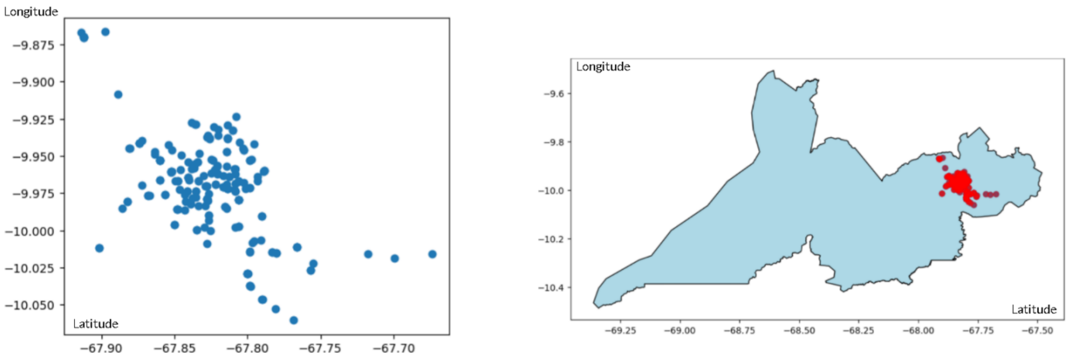


Figure 2: Conceptual model of the organization of geographic entities.

By way of illustration, Figure 3 shows the relationship between spatial data from ERBs. It helps illustrate the argument that integrating data from the nearest stations with their geographic position contributes to the study of improving the architecture of mobile telecommunications networks for the provision of services in the coverage area, with the geographic and geodetic conditions presented in the IBGE data [14]. Figure 3 (a) shows the geometric distance relationship between ERBs in Rio Branco, capital of Acre, and Figure 3 (b) shows the cartographic relationship between ERBs. Therefore, it is possible to visualize the concentration of ERBs in the east of the capital due to a large part of its area being covered by the Amazon Rainforest.



(a)

(b)

Figure 3: (a) Graphical representation of the cartographic distribution of ERBs in the metropolitan area of the East Zone of Rio Branco, and (b) overlay of these ERBs on the cartographic base of the capital.

3.3. Graph of knowledge

Based on the conceptual models presented, a large graph was constructed representing the MOSAICO system ERB data and IBGE data using Neo4J's CYPHER language. The Figure 4 shows part of the graph referring to the structure of ERB 64300. This figure shows the main mobile technologies: GSM (2G), WCDMA (3G), LTE (4G), and NR (5G); the antennas associated with them; equipment that operates their respective antennas; and which antennas are installed in a new project without legacy infrastructure (Greenfield). By looking at the graph, it is possible to understand the structure of how the ERB operates, how its antennas are connected, and how the spectrum of one antenna affects that of another, knowing at what power it operates.

Given its size and scope of representation, viewing the graph in its entirety is not very informative. For this reason, navigation through the graph is done through specific queries written in the CYPHER language.

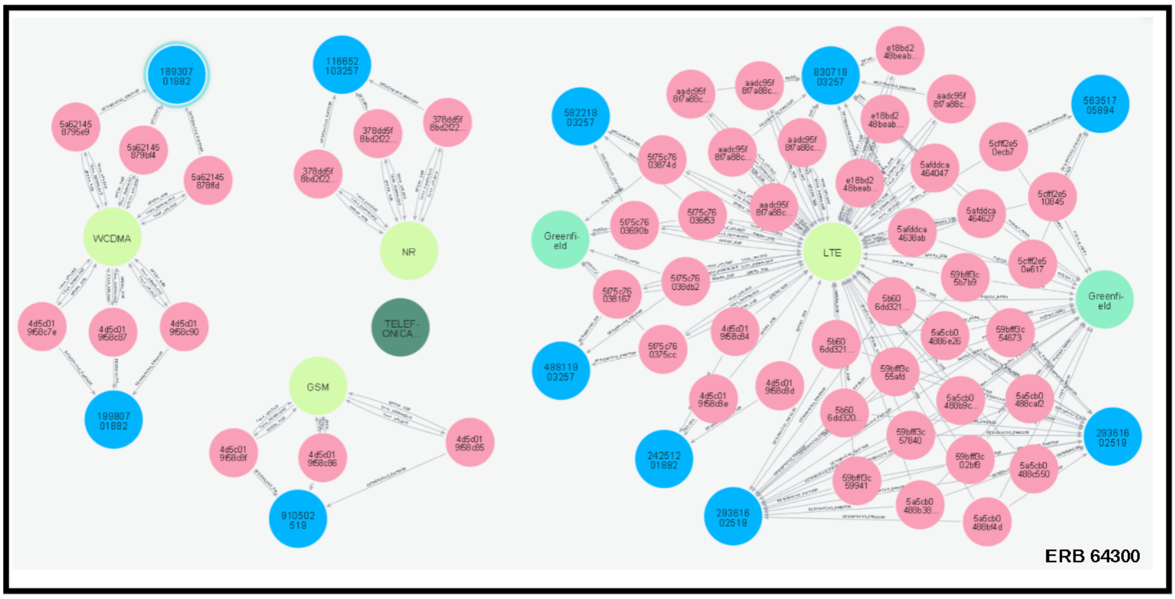


Figure 4: Structure of ERB 64300.

4. Examples of the queries

In order to evaluate the cohesion of the graph and its possible everyday applications, several queries were performed, such as:

- physical and topological structure of ERB 64300 with the number of antennas and their technologies;
- antennas with different physical infrastructure, including the identification number of each antenna and its classification according to the type of sharing;
- Identification of nearby stations and the operators that run them;
- distribution of mobile telecommunications infrastructure by city, state, or region.

The queries performed showed how a knowledge graph can be used to model data and extract structured information, both visually (through graphs) and in JSON and GeoJSON formats. More significantly, the queries highlighted how the semantics of the geographic relationships between ERBs can contribute to the extraction of information about physical and logical connections between infrastructure elements.

5. Conclusion

The challenge of representing spatial data in a manner consistent with its logical and semantic context remains relevant, especially when dealing with georeferenced information. Traditional approaches based on spatial tables often fail to capture the semantic relationships between attributes, limiting the integrated analysis of geographic entities and their interactions.

This work proposed an alternative based on knowledge graphs to integrate spatial and non-spatial databases. Using open data from Anatel and IBGE, it was possible to structure a model that not only represents the distribution of radio base stations in Brazil but also allows for richer analyses of their geoeconomic impact and the interoperability of telecommunications networks (2G, 3G, 4G, and 5G).

The results show that this approach can significantly contribute to public policy decision-making, helping Anatel identify areas requiring infrastructure investments or regulatory adjustments to ensure efficiency and equity in access to telecommunications services.The Future work could explore applying this model in other geographical contexts or sectors, expanding its usefulness for urban and regional planning.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4, Language Tools and Grammarly in order to: Grammar, spelling and plagiarism check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] IBGE, Cidades e estados do brasil, 2022. URL: https://cidades.ibge.gov.br, acesso em: 5 nov. 2024.

[2] IBGE, Malhas territoriais, 2022. URL: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html?edicao=36516&t=downloads, acesso em: 11 out. 2024.

[3] ANATEL, Mosaico - sistema integrado de gestão e controle do espectro, 2016. URL: https://sistemas.anatel.gov.br/se/public/view/b/licenciamento.php?view=licenciamento, acesso em 19 de outubro de 2024.

[4] A. Singhal, Introducing the knowledge graph: Things not strings, 2012. URL: https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html, 16/05/2012.

[5] M. A. Rodriguez, The rdf virtual machine, Knowledge-Based Systems 24 (2011) 890–903. URL: https://www.sciencedirect.com/science/article/pii/S0950705111000860. doi:10.1016/j.knosys.2011.04.004.

[6] J. F. Sowa, Knowledge representation: Logical, philosophical, and computational coundations, Brooks Cole, Pacific Grove, CA, 2000.

[7] I. Basicevic, M. Popovic, N. Cetic, Ontology for ip telephony networks, ICDT 2013 (2013) 9.

[8] Q. Zhou, A. J. Gray, S. McLaughlin, Toco: an ontology for representing hybrid telecommunication networks, in: The Semantic Web: 16th International Conference, ESWC 2019, Portorož, Slovenia, June 2–6, 2019, Proceedings 16, Springer, 2019, pp. 507–522.

[9] Y. Huang, X. You, H. Zhan, S. He, N. Fu, W. Xu, Learning wireless data knowledge graph for green intelligent communications: Methodology and experiments, arXiv preprint arXiv:2404.10365v1 (2024). URL: https://arxiv.org/abs/2404.10365v1. doi:10.48550/ARXIV.2404.10365, submitted April 16, 2024.

[10] F. Aranha, Sistema de informação geográfica: Uma arma estratégica para o database marketing, Fundação Getúlio Vargas 36 (1996) 12–16. URL: https://doi.org/10.1590/s0034-75901996000200003. doi:10.1590/s0034-75901996000200003.

[11] A. N. de Telecomunicações (Anatel), Glossário de termos da anatel: Estação rádio base (erb), 2022. URL: https://informacoes.anatel.gov.br/legislacao/glossario?catid=5&faqid=2902, acesso em 1 de julho de 2025.

[12] Agência Nacional de Telecomunicações (Anatel), Ato nº 9727, de 06 de junho de 2022: Aprova os requisitos técnicos para avaliação da conformidade de equipamentos dos sistemas de acesso fixo sem fio para a prestação do serviço telefônico fixo comutado – stfc, https://informacoes.anatel.gov.br/legislacao/atos-de-certificacao-de-produtos/2022/1681-ato-9727, 2022. Publicado em 8 de julho de 2022; última atualização em 25 de outubro de 2022.

[13] Agência Nacional de Telecomunicações (Anatel), Portaria Anatel nº 2453, de 6 de setembro de 2022: Aprova o procedimento de fiscalização para verificação do cumprimento dos compromissos de abrangência e da Área de cobertura do serviço móvel pessoal (smp), https://informacoes.anatel.gov.br/legislacao/procedimentos-de-fiscalizacao/1724-portaria-2453, 2022. Publicado em 9 de setembro de 2022; última atualização em 26 de novembro de 2024.

[14] D. N. Tavares, Abordagem para análise de dados da infraestrutura de telecomunicação móvel do Brasil baseada em grafo de conhecimento, Master's thesis, Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brasil, 2025.

A Semantic Framework to Compose RESTful Services

Luís Antonio de Almeida Rodriguez[1,*,†], Francisco de Oliveira[1] and
José Maria Parente de Oliveira[1]

[1] *Aeronautics Institute of Technology, São José dos Campos-SP, Brazil*

Abstract

The widespread adoption of RESTful services has led to a fragmented representational ecosystem, where the lack of a formal semantic description model hinders the reliable and automated composition of services. Reliance on non-standardized and non-machine-processable documentation introduces integration inconsistencies in distributed functionalities, requiring manual intervention to resolve ambiguities. This paper proposes an ontology-based semantic orchestration model for the composition of RESTful services, addressing two critical gaps: (i) the absence of a semantic model for describing RESTful service features; and (ii) the lack of semantic matching mechanisms for semantic orchestrations. This model proposes an ontological approach for the semantic orchestration of atomic and composite RESTful services' processes, enabling automated discovery and aggregation. This paper also presents an experimental evaluation within the context of the ICAO's SWIM program, demonstrating the effectiveness of composing RESTful shared services. Moreover, the approach reduces manual integration efforts compared to existing methods. The results suggest that the framework provides a replicable paradigm for large-scale service composition, representing an advance toward sustainable interoperability in heterogeneous distributed systems.

Keywords

RESTful services, Service Composition, OWL-S, Ontologies

1. Introduction

The growing adoption of RESTful services as a dominant architecture for web-based system integration has brought significant advancements in scalability, simplicity, and interoperability over the last decade. This type of service on the Web [1] is typically lightweight, stateless, and aligned with HTTP standards, making it suitable for cloud-native applications, public data platforms, and domain-specific APIs. However, despite their technical advantages, the RESTful paradigm suffers from a profound limitation when it comes to composing multiple services into a coherent workflow [2, 3]. Unlike earlier service-oriented approaches such as SOAP with WSDL and UDDI, RESTful services often lack formal mechanisms for description, discovery, and orchestration, which hampers their reuse and automation across heterogeneous environments.

One of the critical challenges in composing RESTful services lies in their documentation. In most cases, RESTful APIs are described informally through human-readable resources such as HTML pages, Swagger/OpenAPI files, or vendor-specific guides [4]. Although these formats may help developers during manual implementation, they do not allow machine-based reasoning or semantic search [5]. As a result, service integration becomes an error-prone, non-scalable, and context-dependent process that depends heavily on developer interpretation and custom integration logic. This becomes particularly problematic when attempting to aggregate services provided by independent organizations, each adhering to different naming conventions, data structures, and interface assumptions.

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

[†] These authors contributed equally.

✉ rodriguezlaar@gmail.com (L. A. d. A. Rodriguez); francisco.oliveira@ifsp.edu.br (F. d. Oliveira); parente@ita.br (J. M. P. d. Oliveira)

🌐 linkedin.com/in/luÃŋs-antonio-rodriguez-msc-641b3021 (L. A. d. A. Rodriguez); www.francisco.ifsp.edu.br (F. d. Oliveira); https://www.linkedin.com/in/jose-m-parente-de-oliveira-53188618/ (J. M. P. d. Oliveira)

iD 0009-0002-8669-2425 (L. A. d. A. Rodriguez); 0000-0001-5308-7355 (F. d. Oliveira); 0000-0002-7803-1718 (J. M. P. d. Oliveira)

The absence of a unified standard to semantically describe the behavior and capabilities of RESTful services exacerbates the situation [3, 5]. Without a shared vocabulary or formal representation of the functionalities of the service, there is no reliable way to automate the discovery, matching, or composition of services based on their meaning or role in a process. This severely limits the potential for dynamic service orchestration in domains that demand agility, such as meteorology, logistics, healthcare, and air traffic management. For these sectors, the ability to semantically describe and connect service processes is fundamental to achieving interoperability, scalability, and intelligent automation.

This paper addresses these issues by proposing a semantic framework for describing and orchestrating RDF Knowledge Graphs representing RESTful services using ontologies, specifically by extending an OWL-S-based model presented in [5], which extends OWL-S to RESTful architectures. The identified core problem is the absence of an agreed-upon semantic representation for service processes that enables semantic orchestration and reasoning, which will enable automatic composition based on QoS criteria by any API. The proposed solution introduces an ontologically grounded approach for describing RESTful service capabilities, including the associations between atomic and composite processes, with an agreed-upon semantic language that integrates and allows automated extraction of information from the ontologies. A case study based on a meteorological service platform used in aviation (e.g., REDEMET-API) presents the effectiveness and applicability of this model within the broader context of global civil aviation and SWIM (System Wide Information Management).

2. Semantic Standardization in SWIM and the Role of OWL-S

The System Wide Information Management (SWIM) program, coordinated by the International Civil Aviation Organization (ICAO), is a global initiative that standardizes information exchange and fosters interoperability across civil aviation systems [?]. SWIM incorporates data exchange models, governance methodologies, and infrastructure patterns to ensure efficient, secure, and machine-readable distribution of air traffic information. Its primary objective is to enhance the availability, accessibility, and consistency of air navigation data across national and organizational boundaries. By complementing traditional human-to-human communication with structured and automated information flows, SWIM establishes an environment where stakeholders benefit from high-quality semantic data exchange, which is critical for safety, scalability, and international collaboration [?].

The alignment of SWIM with W3C Semantic Web standards is achieved through the OWL-S ontology, which provides the formal structure to represent service capabilities, behaviors, and grounding protocols in a machine-interpretable manner. ICAO and EUROCONTROL documentation emphasize OWL-S as a foundational tool for semantic interoperability, ensuring precise, reusable, and automated service descriptions [?]. However, since OWL-S was originally grounded in SOAP-based infrastructures, extending its principles to RESTful architectures is necessary. This research contributes by adapting OWL-S to REST semantics—incorporating HTTP verbs, URIs, content negotiation, and resource representation types—thus bridging the gap between traditional SOAP/WSDL-based services and modern Web services. This ontology-driven extension enables automated discovery, quality-aware composition, and accurate invocation of RESTful services, fulfilling SWIM's requirements for semantic-native interoperability in heterogeneous and distributed aviation environments.

3. A Semantic Orchestration Model

The goals of this section are the following.

1. To extend the work presented in [5] and create a semantic model to support the orchestration of service process relationships to automate the composition of services by APIs. Building upon the OWL-S extension developed in this paper, where RESTful services' processes have an ontologically defined taxonomy using 'Atomic' and 'Composite' classifications, this work involved classifying those implemented SOA-unit service processes and building an RDF Knowledge Graph by associating them, creating a RDF Knowledge Graph.

2. Refactor the software proposed in [5] and develop a new functionality to perform automated dynamic service composition, driven by the common web user's own choices. Specifically, this functionality leverages the semantically enriched processes already orchestrated in the model and links them with the REDEMET-API RESTful services, generalizing the composition coded with Python to accommodate service descriptions with different abstraction levels.

The orchestration logic, grounded in OWL-S' ontological compatibility, enables any API to integrate diverse services into cohesive workflows without requiring manual intervention. This enhancement addresses the broader vision of democratizing service composition, allowing non-expert users to trigger complex sets of services by specifying 'Quality of Service Level' ranks for each one. In doing so, this paper's proposal reinforces the practical value of semantic description environments in real-world web architectures and extends the OWL-S adaptation presented in [5] toward multi-provider interoperability for intelligent compositions.

The central point of this section is the *Process* ontology, exactly as it is presented in [5]. This ontology, within the extension of OWL-S proposed in this paper, serves as a foundational component to describe the behavioral dynamics of services on the Web. Its internal structure is systematically organized according to several defining criteria that establish the functional and operational semantics of service processes. These criteria facilitate the semantic search, and the specific Atomic-Composite association facilitates service composition. Central to this ontology are the Inputs, Outputs, Preconditions, and Effects (IOPE), which provide a formal specification of a service's capability, allowing for precise semantic matching during service discovery. Precise search and discovery of services are essential prerequisites for service composition; finding the best services enables the creation of an optimal composition.

Martin [6] comments that Atomic processes represent indivisible actions with direct invocations, while Composite processes, on the other hand, define complex associations through atomic or other composite processes. The *Process* ontology's ontological representation enforces a strict taxonomic distinction whereby every individual of the **Process** class must also be instantiated as a member of **AtomicProcess**. Figure 1 shows that some of them are instances of the **CompositeProcess** class, in full compliance with the foundational specifications of OWL-S. This mandatory classification is formally realized through two canonical inverse properties, as depicted in the modeling presented in [5]. Figure 1 illustrates that:

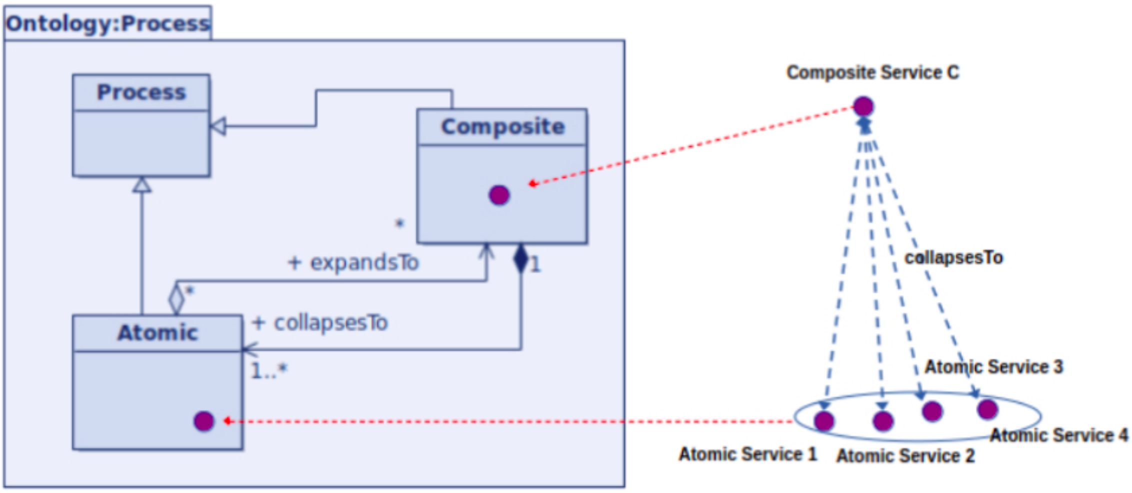


Figure 1: Orchestration's Model of Process entities.

- collapsesTo - A decomposition linkage: **CompositeProcess** → **AtomicProcess**

- expandsTo - An ’Aggregation’: **AtomicProcess → CompositeProcess**

This dual-property framework ensures both structural integrity and semantic consistency in process decomposition, thereby enabling rigorous reasoning over service composition orchestrations such as this one. Using these properties, it is possible to select all instances of services already implemented and insert their ’process entities’ into the existing OWL classes: **Atomic** and **Composite**, in the *Process* ontology. After that, associations among composite and atomic entities are necessary using object properties already defined in this ontology (e.g., collapsesTo and expandsTo). Why classify them all as atomic and some as composite? By doing this, it is possible to establish two types of composition:

- A composite service composed only of atomic services, and
- A composite service composed of atomic and other composite services.

The formalization of service processes classification and maximization was given considering the OWL-S Compliant Knowledge Graph. According to [5], all instances of service processes within the developed knowledge base have been systematically classified as members of the **AtomicProcess** class, and some were classified in the **CompositeProcess** class. This strict taxonomic enforcement ensures adherence to the foundational semantics of OWL-S and a new characteristic: the ability to create composite services composed of *other composite services*.

To establish machine-interpretable semantic relationships between processes, each instance of a process from the semantic registry built in [5] has been explicitly associated with the canonical object properties collapsesTo and expandsTo, thereby formalizing hierarchical dependencies within service compositions:

- AtomicProcess individuals represent an indivisible service set of operations, with no further decomposition permitted under the axiomatic constraints of the ontology (e.g., even for ’composite services’ classified as such).
- CompositeProcess individuals are structured aggregations of subprocesses, recursively defined via the collapsesTo property, which links a composite service to its constituent atomic or nested composite processes.

The resultant RDF knowledge graph encodes these relationships as semantically described triples, conforming to OWL-S specifications while enabling advanced querying capabilities. Through SPARQL queries, the graph can be interrogated to:

- Retrieve atomic services meeting specific functional criteria (e.g., inputs/outputs, preconditions);
- Decompose composite services into their procedural workflows, exposing nested execution logic;
- Validate process compositions by reasoning over property constraints.

The structured representation presented in Figure 1 collaborates with more criteria for the discovery of semantic services and the analysis of automated composition, as the formal semantics of the graph permit inference about the compatibility of processes and behavioral equivalence. It is possible to mention some consequences of applying our model proposed in Figure 1 to the resulting RDF Knowledge Graph:

- Establishment of a hierarchical process topology
- Creation of an explicit definition of the service composition paths through relationships among nodes, represented by Triples
- A layer of descriptions made by semantic machine-readable composition patterns

By formalizing process hierarchies, constraints, and operational logic, the Process Ontology ensures that services can be dynamically discovered, composed, and invoked in a semantically consistent manner.

4. Case Study

RESTful aeronautical service development communities encounter systemic interoperability barriers due to representational fragmentation: a divergence in service description models [5]. This causes a huge development effort to build APIs able to use RESTful shared services on the Web.

4.1. Motivation

This study addresses the urgent need to realize SWIM's vision of seamless and standardized air traffic management (ATM) data exchange between the ICAO member states, aiming to improve the ability to use shared RESTful services [7] to create more complex services. SWIM's proposed semantic enrichment of service descriptions transforms interoperability from a legacy concept into a natively standardized communication paradigm, enabling a globally harmonized aviation data interoperability. Building on existing frameworks, this paper aims to bridge these gaps, ensuring deterministic composition made by software through semantically enriched service descriptions.

4.2. Orchestrating and Composing Complex RESTful Services

The article by Rodriguez and Parente de Oliveira [5] presents a rigorous semantic model designed to enable the composition of RESTful services through ontologically grounded orchestration. Drawing on the foundational architecture of OWL-S, the model presents a dual-level abstraction also presented in the original OWL-S: atomic and composite processes. Each SOA-unit service registered in this semantic registry has its process entity semantically annotated and encapsulated within a class called Atomic process, which is connected, in the RDF Knowledge Graph, with entities that describe its capabilities, such as inputs, outputs, preconditions, and effects (IOPE).

Composite processes are structured as orchestrations of atomic processes made to define the logical and temporal relations between service calls [8, 5]. This semantic representation enables reasoning engines to identify, match, and bind services dynamically, thereby automating the composition of services into coherent workflows. To instantiate the model presented in [5], the authors applied it in a case study involving the semantic representation and orchestration of 23 SOA unit services from the original model. The propagation of the orchestration of SOA-unit services will reach 17 RESTful services provided by REDEMET-API [4], and three other RESTful services as part of a provider specifically created in the registry for this paper, the **SWIM Provider**. The SOA-unit services' process entities were first formalized as an atomic process, where the semantic annotations create an atomic foundation for serving several composite processes.

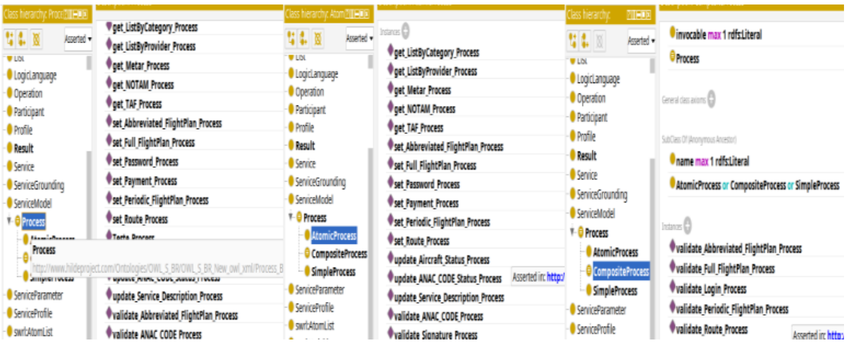


Figure 2: Process entities orchestrated in Protegé.

Figure 1 presents the proposed orchestration. This hierarchical structure enables the expression of increasingly complex service behaviors while preserving semantic clarity. This layered composition model demonstrates the flexibility of OWL-S in describing realistic service orchestration scenarios in the RESTful paradigm. There are no cyclic representations in the RDF Knowledge Graph. The functionality is granted through an accurate set of associations between all composite and atomic processes.

The semantic orchestration presented in this paper was implemented in the *Process* ontology, such as in [5] using Protégé 5.5. The procedure begins with the import of the OWL-S ontology into a new project. The entity of the process of each generic service was created in the **ServiceModel** class as an instance of the `AtomicProcess` class. The HTTP semantics, including the type of the method and the endpoint URI, were encoded through custom data properties or extended using existing ontologies such as *Aerodrome*, *Country*, and *RESTfulExtension*.

Subsequently, composite processes were created using the `CompositeProcess` class, and their execution flow was defined via semantic relationships between the processes using ***Object Properties*** already existing in the previous model [5]. This ultimately results in a machine-interpretable semantic representation that captures the logic of SOA-unit composite services and propagates to the constraints and sequencing of the REDEMET-API RESTful services network. This case study implementation effort produced a machine-interpretable semantic infrastructure that not only encapsulates the logic of composite SOA-unit services but also propagates the predefined features about process sequencing across the REDEMET-API meteorological RESTful services network.

The case study practical outcomes are multifold: for the developers community, it offers a reusable and extensible description model for service composition's orchestration; for global civil aviation, the model aligns with SWIM's vision of semantic interoperability and dynamic information sharing; for end users and companies, it enables intelligent service discovery and automated interaction, thus reducing integration costs, enhancing scalability, and paving the way for a new generation of semantically aware RESTful ecosystems.

The orchestration model proposed and implemented in this study provides concrete generic-domain benefits by enabling a semantic composition framework that propagates from generic SOA-units to specific RESTful services. By providing a formal semantic description, developers can automate essential events of the composition pipeline, replacing ad hoc scripting and manual configuration with logically inferred orchestration. By offering semantic transparency and structural formalization, the orchestration mechanism allows heterogeneous provider services to be composed into reliable workflows without human intervention, enhancing situational awareness, operational efficiency, and safety. Beyond aviation, the approach serves as a generic technological contribution to the field of service composition by demonstrating that OWL-S-based semantic models can be successfully adapted to RESTful paradigms.

5. Conclusions

5.1. Contributions

A key contribution is the definition of a novel orchestration format, grounded in OWL-S and enriched with precise object properties (e.g., expandsTo, collapsesTo), which allows developers to express composition relationships among services with reduced syntactic ambiguity. It reduces the cognitive and technical effort required to construct complex service integrations, exactly as evaluated in [5], replacing heuristic-based composition methods with logic-based reasoning. Moreover, by implementing and validating the model over a real-world API—REDEMET, the paper contributes a reusable, standards-aligned methodology for modeling and composing RESTful services in domains requiring high semantic precision, such as civil aviation under the SWIM framework.

Declaration on Generative AI

During the preparation of this work, the author(s) used Deep Seek as a tutor for English Grammar Correction.

References

[1] R. T. Fielding, Architectural Styles and the Design of Network-based Software Architectures, Ph.d. dissertation, University of California, Irvine, Irvine, CA, USA, 2000. URL: https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf.

[2] Y. Gamha, A framework for rest services discovery and composition, Service Oriented Computing and Applications 17 (2023) 259–275. doi:10.1007/s11761-023-00376-6.

[3] M. Bennara, M. Mrissa, Y. Amghar, An approach for composing restful linked services on the web, in: Proceedings of the 23rd International Conference on World Wide Web, Seoul, South Korea, 2014. doi:10.1145/2567948.2579222, hal-01212722v2.

[4] REDEMET-API, Brazilian Aeronautical Provider, https://ajuda.decea.mil.br/base-de-conhecimento/api-redemet-o-que-e/, 2025. Accessed: Apr. 16, 2025.

[5] L. A. A. Rodriguez, J. M. P. de Oliveira, A semantic model to describe restful services, IEEE Access 13 (2025) 72402–72426. doi:10.1109/ACCESS.2025.3562503.

[6] D. Martin, Bringing semantics to web services: The owl-s approach, IEEE Intelligent Systems 22 (2007) 72–81. doi:10.1109/MIS.2007.100.

[7] EUROCONTROL, Swim technical infrastructure and governance: Semantic interoperability guidelines, 2019.

[8] OWL-S Coalition, Owl-s: Semantic markup for web services, version 1.2, https://www.daml.org/services/owl-s/1.2/, 2006. Released by the OWL-S Coalition, including members from SRI International, Carnegie Mellon University, BBN Technologies, and University of Maryland.

Semantic Mapping of Bibliographic Models in National Libraries

Felipe Augusto Arakaki[1,†], Ana Carolina Simionato Arakaki[1,2,*,†] and Ana Carolina Novaes de Mendonça[3,†]

[1] *Faculty of Information Science, University of Brasilia, Campus Darcy Ribeiro - DF, 70297-400 , Brasilia, Brazil*
[2] *Graduate program in Information Science, Federal University of São Carlos (UFSCar), São Carlos, Brazil*
[3] *Brazilian Institute of Science and Technology, SAUS Q 5, L 6, Bl H, Brasília, DF, Brazil*

Abstract

This paper presents a comparative analysis of the data and metadata models used by national libraries that publish their collections following the principles of Linked Open Data (LOD). Using the Crosswalk method, the study mapped the classes and properties adopted by different institutions, identifying conceptual convergences, terminological variations, and distinct modeling strategies. The results reveal heterogeneous practices, ranging from simplified models to robust, ontology-based structures. Despite shared core entities such as Work and Person, semantic differences remain regarding the scope and level of detail. The findings support the development of semantic alignment strategies and shared vocabularies to improve interoperability.

Keywords
Linked Data, Semantic Web, Metadata, Ontology, Library.

1. Introduction

The provision of structured bibliographic data on the Web, in an open and connected manner, represents one of the main contemporary challenges faced by libraries. Although libraries have traditionally adopted well-established representation standards, such as Machine-Readable Cataloging (MARC 21) developed by the Library of Congress in the United States, many of these models were originally designed for the construction of printed and centralized catalogs, which hinders their integration into digital ecosystems guided by the principles of the Semantic Web and Linked Open Data (LOD). In this context, transforming bibliographic records into connected data requires technological adjustments and a theoretical-methodological revision of the approaches to data modeling, description, and interoperability.

The publication of linked open data requires the use of persistent identifiers, standardized vocabularies, a consistent semantic structure, and machine-readable formats. Various international organizations, such as the World Wide Web Consortium (W3C), have promoted best practices for publishing linked data, aiming to build an interoperable, accessible, and reusable Web of Data. However, in the bibliographic domain, the heterogeneity of data models adopted by different institutions, as well as the scarcity of initiatives that effectively publish their data in accordance with these principles, reveal the complexity of transitioning from legacy systems to LOD-oriented architectures.

This article investigates the issue of structuring bibliographic data in national libraries based on the principles of LOD. The aim is to identify the data and metadata models used by libraries that

[*] *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.*
[1*] Corresponding author.
[†] These authors contributed equally.

✉ felipe.arakaki@unb.br (F. A. Arakaki); ana.arakaki@unb.br (A. C. S. Arakaki); anamendonca@ibict.br (A. C. N. Mendonça)
iD 0000-0002-3983-2563 (F. A. Arakaki); 0000-0002-0140-9110 (A. C. S. Arakaki); 0009-0004-0285-9932 (A. C. N. Mendonça)

already publish their collections as connected data, analyzing their classes, properties, and representation strategies. The research focuses on the systematization of existing models through the application of the Crosswalk method, which enables comparative analysis and semantic mapping across the different standards in use.

The relevance of this study lies in providing support for the harmonization of bibliographic data at the international level, promoting greater interoperability, visibility, and reuse of national collections. By identifying convergences and divergences among the models in use, the study seeks to understand how libraries have been adapting their descriptive structures to the context of linked data, and which methodological paths can be followed by institutions that have not yet adopted this publication model.

2. Related works

The proposal for structuring and publishing bibliographic data on the Web is situated within the broader context of the Semantic Web and the principles of LOD [7]. The Semantic Web aims to assign meaning to data available on the internet so that it can be understood and processed by computational agents, enabling automated integration, discovery, and reuse of information. LOD, in turn, refers to a set of best practices for publishing structured data that are interlinked by URIs and described using interoperable standards such as Resource Description Framework (RDF), Web Ontology Language (OWL), and SPARQL Protocol and Resource Description Framework Query Language (SPARQL) [8,9].

In the field of Library and Information Science, the adoption of these principles requires the adaptation of traditional tools that support information representation such as MARC21 and the Universal Machine-Readable Cataloging (UNIMARC), developed by the International Federation of Library Associations and Institutions (IFLA) toward an ontology-driven structure. Initiatives such as BIBFRAME, developed by the Library of Congress, represent concrete efforts in this direction by proposing a model based on entities such as Work, Instance, and Item, aligned with controlled vocabularies and Semantic Web standards. Nevertheless, the diversity of models adopted by national libraries in different countries presents challenges for interoperability and semantic mapping.

It is important to highlight the risks and benefits of simplifying schemas for interoperability purposes: while such simplification enhances machine readability, it may lead to the loss of descriptive nuances, requiring critical decisions about which characteristics of entities should be preserved [1]. In this context, ontologies play a central role as tools for conceptual structuring, enabling the explicit representation of classes, properties, and relationships between domain entities, thus promoting consistency and reuse across heterogeneous systems [10].

The present research is therefore grounded in approaches that integrate the Semantic Web, Linked Data, and metadata schema mapping, aiming to understand the models adopted by national libraries and to propose pathways for interoperability and the publication of bibliographic data as LOD.

From a methodological standpoint, several studies propose strategies for mapping between metadata schemas, highlighting the importance of preserving semantic integrity during the conversion process [2]. Additionally, the challenges of achieving semantic interoperability are underscored by the diversity of elements and structural variations adopted by different information communities [1].

The literature also highlights the central role of ontologies in the representation of bibliographic data in digital environments. Studies provide guidelines for defining classes, properties, and relationships between entities [11], while others discuss the impacts of schema simplification on metadata quality and reusability [12].

3. Methodology

This research is characterized as qualitative, exploratory and theoretical, using the Crosswalk method, proposed by the National Information Standards Organization (NISO) in 1999, for data analysis. The Crosswalk method enables interoperability between systems that use heterogeneous metadata standards. According to NISO [1] "Crosswalks provide the ability to make the content of elements defined in a metadata standard available to communities using related metadata standards".

The Crosswalk method consists of four stages: harmonization, semantic map, element-to-element mapping, and hierarchy, object and logical view. To carry out the crosswalk, Chan and Zeng [2] highlight two approaches: "absolute crosswalking" and "relative crosswalking". Absolute crosswalking refers to the exact correspondence between metadata, while relative crosswalking is used to minimize information loss by matching elements of a source schema to at least one element of a target schema.

The crosswalk process can face equivalence difficulties, such as one-to-one, one-to-many, many-to-one and one-to-one. However, it was decided to carry out a general mapping, checking the compatibility of the classes and properties proposed by each of the institutions analyzed. The crosswalk was done individually for each metadata standard, followed by the creation of a table with the crosswalk of all the standards analyzed, establishing a general overview of the mapping.

In this context, a manual crosswalk was developed to map the metadata elements used by national libraries whose catalogs share connected open data. The task involved a detailed semantic analysis of each metadata element and its contextual application. Four specialists conducted this work based on the official documentation provided by each institution.

The selection of libraries was guided by a previous survey [3] that identified eleven national libraries publishing linked open data. However, among these eleven institutions, only seven were found to explicitly provide documentation regarding their data and metadata models, which enabled their comparative analysis within the scope of this study. The libraries analyzed were: the National Library of Spain (BNE), the National Library of France (BnF), the German National Library (DNB), the Finnish National Bibliography (FENNICA), the Royal Library of the Netherlands (KB), the Library of Congress (LC), and the National Library of Medicine (NIH).

4. Results and discussion

Based on the results in Table 2 and the specialized literature, it is possible to discuss the diversity of approaches adopted by National Libraries in modeling their bibliographic data in Linked Open Data (LOD) environments. It should be noted that not all classes and proprieties identified in the models are detailed in the text, but only a selection.

The Library of Congress, represented through the BIBFRAME model, was developed to replace the MARC21 format and constitutes an entity-based structure aligned with the principles of the Semantic Web. Among the institutions analyzed, it was the library that showed the highest degree of correspondence with the other models, particularly through the adoption of core classes such as Work and Person. This indicates a baseline convergence around essential elements of bibliographic description. The BIBFRAME model currently comprises 208 classes and 148 properties, which reflects a robust and detailed semantic framework. The observed overlap suggests partial alignment between the models in use, although differences persist regarding semantic granularity and the inclusion of additional entities such as Expression, Contribution, and Dataset, which vary depending on institutional contexts and implementation strategies.

The National Library of Spain (BNE) The National Library of Spain (BNE) adopts the IFLA Library Reference Model (LRM) as the conceptual basis of its bibliographic data model. Its current structure includes 8 classes and 225 properties, reflecting a semantically rich approach oriented toward interoperability and reuse of bibliographic data [4].

Table 2
Mapping of entity classes in National Libraries with Open Data catalogs

BIBFRAME	BNE	BNF	DNB	FENNICA	KB	NIH
Agent/Person	Person	Person	Persons/ Entities	Person	-	Agent
Collection	-	-	-	-	-	Collection
-	-	Concept	-	Concept	-	-
Contribution	-	-	-	-	-	Contribution
Dataset	-	-	-	-	Dataset	Dataset
Item	Item	Item	-	-	-	-
Organization	Corporate Body	Collective agent	-	Organization	-	Agent
Place	-	Place	-	Place	-	-
Work	Work	Work	-	Work	-	-

The Bibliothèque Nationale de France (BnF) adopts a limited set of entity classes Person, Item, Place, and Work reflecting a more selective approach to semantic modeling. The main entities are grounded in the IFLA Library Reference Model (LRM). The BnF integrates different metadata standards such as Internal Format of the National Library (InterMARC)[5].

Among the national libraries analyzed, the Deutsche Nationalbibliothek (DNB) adopts a data model that groups entities under the broader class Persons/Entities, encompassing Person, Concept. The Finnish National Bibliography (FENNICA) [6] includes the classes Person, Organization, Place, Concept, and Work, which demonstrate a concern for semantic granularity and alignment with Linked Data principles.

In contrast, the Koninklijke Bibliotheek (KB) displays 7 classes and 50 properties in the mapped data. With emphasis on Dataset and Work, the limited appearance of other classes like Person or Organization may reflect a partial publication strategy or a modeling focus on specific types of bibliographic resources. This configuration constrains broader analysis of its ontological infrastructure. Lastly, the National Library of Medicine (NIH) adopts a domain-specific model oriented toward scientific and biomedical information. The mapped entities include Agent, Collection, Contribution, and Dataset, revealing a strong emphasis on collaborative authorship and data sharing.

The class mapping conducted offers valuable insight into how different national libraries and bibliographic frameworks structure knowledge and resources. It reveals not only a shared foundation based on core concepts, but also the distinctive modeling choices made by each institution to address its specific informational and operational contexts. The role of reference models such as IFLA-LRM is particularly significant, as it establishes a common baseline of entities that facilitates interoperability across diverse vocabularies and systems.

To further understand the modeling strategies adopted by these national libraries, it is essential to complement the analysis of entity classes with an examination of the properties employed. The analysis of the properties used in open linked data catalogs reveals not only differences in the granularity and descriptive scope adopted by each library but also varying levels of adherence to Semantic Web best practices. Properties such as hasPart, ISBN, and ISSN are widely present across the models analyzed, suggesting a common core of elements focused on the identification and structuring of bibliographic relationships.

However, in the BIBFRAME model used by the Library of Congress, both ISBN and ISSN are treated as classes rather than properties. This means that these identifiers are modeled as autonomous entities with their own attributes and relationships, in line with the entity-oriented approach proposed by the model. This modeling choice reflects an effort to enhance semantic flexibility, allowing, for instance, different editions or versions of a publication to be associated with multiple identifiers while maintaining consistency across complex datasets. This conceptual distinction is critical, as it directly affects interoperability between models and the way data can be linked to other datasets on the Web of Data.

Another relevant aspect is the use of the Description property, present in the catalogs of FENNICA, KB, and NIH. This property, often associated with widely adopted vocabularies such as Dublin Core, indicates a concern with both human-readable records and semantic indexing. Meanwhile, the Date property, used by BNF, DNB, and FENNICA, underscores the importance of temporal elements in bibliographic control, which are essential for organizing editions, versions, and publication events.

Lastly, the hasPart property, recurring in all models analyzed except BIBFRAME (where the relationship is represented differently), demonstrates a broad recognition of the importance of hierarchical and compositional relationships in bibliographic records. This property is crucial for representing collections, volumes, chapters, or any composite structure of works, reinforcing the need for descriptive mechanisms that capture complex relationships among resources.

Taken together, these findings show that, although there is a minimal set of shared properties, modeling decisions such as elevating identifiers to entities or emphasizing certain descriptive properties vary according to technical and institutional contexts. This highlights the need for semantic mapping and alignment strategies to enable full interoperability between catalogs structured as Linked Open Data.

5. Conclusion

This study presented a comparative analysis of the data and metadata models used by national libraries that publish their collections according to Linked Open Data (LOD) principles. Based on the Crosswalk method, it was possible to map the classes and properties adopted by different institutions, identifying conceptual convergences, terminological variations, and specific strategies for structuring bibliographic data.

The results reveal a heterogeneous landscape in the models employed, ranging from simplified approaches, such as those of the Bibliothèque Nationale de France and the Biblioteca Nacional de España, to more complex structures like the Deutsche Nationalbibliothek. Amid this diversity, there is a growing trend toward the adoption of the BIBFRAME model, particularly among Anglophone libraries such as the Library of Congress, which indicates consistent efforts toward standardization and semantic interoperability on an international scale.

The mapping conducted revealed that, although many libraries share core classes and properties such as Work, Instance, Item, Agent, and Concept relevant semantic variations persist regarding their application, scope, and level of detail.

One of the limitations identified is the absence of formal validation of the mappings, which could be addressed in future studies with the support of specific ontology alignment tools. Despite these limitations, the findings contribute meaningfully to ongoing discussions about bibliographic data interoperability in contexts driven by the Semantic Web. The study provides concrete support for the development of shared vocabularies and harmonization strategies among bibliographic systems, assisting in national collection modernization and open access initiatives.

For future developments, it is recommended to create a unified vocabulary that can serve as a foundation for interoperability between institutions, supported by ontologies as central tools for semantic alignment across heterogeneous descriptive standards. Additionally, the use of artificial intelligence techniques such as machine learning and semantic inference offers promising potential to enhance the processes of discovery, reconciliation, and reuse of information across libraries, broadening the reach and effectiveness of data integration on the Web.

It can be concluded that harmonizing bibliographic data and metadata models is a fundamental step toward building a more connected, accessible, and standards-driven information ecosystem, significantly contributing to the strengthening of national bibliographic heritage in the era of linked data.

Acknowledgements

Acknowledgements of support from the National Council for Scientific and Technological Development (CNPq) for the projects: Linked data publishing in libraries: theoretical-methodological proposal for SIBISC, CNPq Universal n° 409407/2021-6, Connected authority data for libraries: theoretical and methodological proposal for SIBISC, CNPq Universal n° 421178/2023-0 and the Brazilian Institute of Information in Science and Technology (Ibict).

Declaration on Generative AI

The AI was used to evaluate the text and assist in the drafting of some paragraphs.

References

[1] Pierre, M. S., & LaPlant, W. P, Issues in crosswalking content metadata standards. [S.l.]: NISO Baltimore, Maryland, USA 2000.

[2] L. M. Chan and M. L. Zeng, Metadata interoperability and standardization: A study of methodology part I. D-Lib Magazine, 12(6), 1082–9873, 2006.

[3] A. F. de. Jesus, Recomendações teórico-metodológicas para a publicação de dados bibliográficos abertos e conectados (Dissertação de mestrado, Universidade Federal de São Carlos, UFSCar, São Carlos, SP), 2021. https://repositorio.ufscar.br/handle/ufscar/14228.

[4] Biblioteca Nacional de España. Ontología BNE (Rev. 2.0). 2020. Biblioteca Nacional de España. https://datos.bne.es/def/index-es.html.

[5] Bibliothèque nationale de France. (n.d.). Semantic Web and Data Model. Bibliothèque nationale de France. https://data.bnf.fr/semanticweb.

[6] Fennica. Fennica RDF data model, 2019. https://www.kiwi.fi/display/Datacatalog/Fennica+RDF+data+model.

[7] PomerantzM Metadata. USA: The MIT press essential knowledge series. 2015.

[8] T. Berners-Lee, Linked data: Design issues. 2006. https://www.w3.org/DesignIssues/LinkedData.html.

[9] B. Hyland, G. Atemezing, and B. Villazón-Terrazas, Best practices for publishing Linked Data, 2014. W3C Working Group Note. Disponível em https://www.w3.org/TR/ld-bp/.

[10] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web. In Proceedings of the 17th International Conference on World Wide Web (pp. 1265–1266). [*S.l.*]: ACM.

[11] S. Van Hooland and R. Verborgh, Linked Data for Libraries, Archives and Museums: How to clean, link and publish your metadata. Facet Publishing. 2014

[12] N. F. Noy and D. L. McGuinness, Ontology development 101: A guide to creating your first ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05, 2001.

Providing Semantics to Object-Centric Event Logs (OCEL) from Process Mining using BFO

Evellin Cardoso[1]

[1]Federal University of Goias, Goias, Brazil

Abstract

Business Process Management (BPM) is a discipline encompassing a wide array of methods, techniques, and tools derived from both Information Technology and Management Sciences to manage business processes within organizations. This paper addresses a critical BPM challenge, specifically concerning the semantic enrichment of process data. We propose a novel framework designed to augment Object-Centric Event Logs (OCEL), a recent advancement in Process Mining, with the robust ontological foundation provided by the Basic Formal Ontology (BFO). By integrating OCEL data with BFO, our framework seeks to enrich process mining artifacts with deeper semantic meaning, thereby enabling more sophisticated analysis, more comprehensive understanding of business processes, also opening the possibility to obtain richer insights from Process Mining algorithms.

Keywords

Ontology, Basic Foundational Ontology (BFO), Process Mining, Data, Knowledge-Augmented Business Process Management (BPM)

1. Introduction

The Business Process Management (BPM) discipline provides a structured approach to understanding, optimizing, and controlling the business processes that operate in enterprises [1]. Central to BPM is the concept of a business process, which can be understood as a coordinated set of activities performed to handle a specific case or process instance [1]. The BPM lifecycle typically involves several phases: (i) the (re)design phase, in which process models are elaborated; (ii) the configuration phase, where these models are implemented within a Business Process Management Systems (BPMS), and (iii) the execution phase, during which numerous process instances are created, with sequences of activities reflecting the evolution of each instance. Traditionally, process analysis has heavily relied on classical event logs, which capture activities identified by a single case ID.

The field of Process Mining, which systematically utilizes event data to improve business processes [2] witnessed the emergence of Object-Centric Process Mining (OCPM) techniques. OCPM introduces Object-Centric Event Logs (OCEL) [3], a paradigm shift in how event data is organized. Unlike classical event logs that are case-centric, OCELs are structured in relation to the various data objects involved in a business process, rather than solely focusing on a single case notion. While OCEL offers a more granular and interconnected perspective on data, it still predominantly operates within the procedural sphere, focusing on timestamped activities and how they update data object attributes, without explicitly incorporating the broader enterprise knowledge structure.

To tackle this gap between knowledge and data, this paper addresses the challenge of augmenting the event data stored in OCEL logs with process and domain knowledge. To achieve this, we propose a novel framework that integrates OCEL data with ontologies, specifically leveraging the Basic Formal Ontology (BFO) [4]. Ontologies, as a cornerstone of Semantic Web technologies, provide a common vocabulary for representing knowledge and information across heterogeneous resources and applications, thereby promoting data integration and interoperability. BFO, an ISO standard since 2002, is chosen for its

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

*Corresponding author.
evellin@ufg.br,evellinc@gmail.com (E. Cardoso)
0000-0001-6242-662X (E. Cardoso)

realist stance to capture general features of reality. By bridging OCEL with BFO, our framework aims to provide a semantically rich representation of business processes, enabling a deeper level of analysis and insight that transcends the limitations of purely procedural data. This integration thereby paves the way for more intelligent and robust process management solutions.

This paper is structured as follows. Section 2.1 introduces the BPM discipline and the Basic Foundational Ontology (BFO). Section 3 presents the OntoOCEL framework that augments data from OCEL logs with BFO concepts. Section 5 concludes the paper, outlining directions for future work.

2. Baseline

2.1. Business Process Management (BPM) and Process Mining

Business Process Management (BPM) is the discipline that includes methods, techniques and tools from Information Technology and Management Sciences to support the design, analysis, configuration, execution and monitoring of business processes [1].

The notion of business process is central to BPM. A *business process* consists of a set of activities that are executed to handle a *case* (a process instance) [1]. To manage business processes, BPM conducts its efforts along different phases of the *BPM lifecycle* [1]. In the (re)design phase, a process model is elaborated. This model is subsequently configured/implemented in a business process management system (BPMS) within the configuration phase. Within the execution phase, processes instances are created multiple times, with sequences of activities corresponding to the evolution of each case.

Taking advantage of this event data and process models, *Process Mining* is the discipline that systematically uses event data to improve business processes [2]. Traditional process mining techniques rely on classical event-logs which are based on the assumption of a single case notion, with events referring to exactly this case. More recently, Object-Centric Process Mining (OCPM) introduces Object Centric Event Logs (OCEL) which are organized in relation to the data objects present in the business, rather than in relation to the cases as a sequence of events. Although this novel way of organizing events in OCEL reflects a more seamless way to look into the data, which is closer to the process as it actually is, the data analyzed by OCPM algorithms still heavily remain in the processual sphere of the business.

The fact that only procedural knowledge is tracked in BPMS system has been already identified as one of the major unresolved issues in BPM [5]. In this work, authors highlight that event logs lack domain-specific and commonsense knowledge, which make them often suffering from noise and incompleteness, thus impacting the outcomes and insights provided by traditional process mining algorithms. To tackle this problem, we augment the event data from OCEL logs with process and domain knowledge in this paper.

2.2. Ontologies and Semantic Technologies

Recently, Semantic Web technologies started gaining increasing attention for their ability to promote heterogeneous data integration and interoperability. Ontologies provide a common vocabulary for representing knowledge and information across heterogeneous resources and applications.

To support such interoperable environment, different types of ontologies exist [4]. *Upper-level ontologies* provide a highly general vocabulary of categories and relations regardless of domain, while *domain ontologies* covers a basic set of universal categories from particular scientific domains.

There exists a number of upper-level ontologies already in place, such as DOLCE, UFO, GFO, BFO, etc [4]. The Basic Foundational Ontology (BFO) is chosen in this paper to promote semantic interoperability because it has been introduced as an ISO standard since 2002, including BFO's ISO 21838-2 specification being axiomatized in First-Order Logic, OWL 2, and CLIF [6].

BFO top-level ontology has been designed with a realist instance in mind. Its main goal is to use ontologies to represent the knowledge acquired as a result of scientific efforts. Being realist means that such knowledge captures general features of reality as general theories (i.e. generalizations and laws of science), rather than particular facts. To make such distinction, BFO presents some fundamental categories as described below:

Universal, **Particular** and **Defined Classes.** Universals are mind-independent entities that can be repeatedly instantiated across time and space, with an indefinite number of particulars, while particulars are individual entities, restricted to specific places. They instantiate universals, but cannot be instantiated. Defined classes are general terms used in science to refer to particular individuals in reality (e.g., medical doctor, dog) that have no corresponding universals [4]. In this way, medical doctor would simple be an instance of person (an Universal) that bear role of medical doctor.

Relations: to relate these entities, BFO introduces three basic relations: (i) universal-universal, (ii) universal-particular and (iii) particular-particular. While universal-universal relations connect subtypes to parent types (*IS A* relation), universal-particular relations relate the instances (particulars) to the universals in which they fall (*instance of*).

BFO is structured in terms of two disjoint hierarchies of universals, depending on how particulars relate to time. Continuants and occurrents are the roots of each branch:

Continuant and **Occurrent:** Continuants endure through time, retaining their identity, fully existing at any time they exist. Examples include a house, an apple, color of an orange. In contrast, occurrents unfold over time, being composed of temporal parts. Examples include a talk, a race, the history of Brazil, a period of time in which the sun rises, people attending a meeting.

Process and **Temporal instant:** A process is an occurrent composed by some temporal proper parts, a time interval $I = [t_i, ..., t_f]$ in which a *material entity* is as *participant*, while a temporal instant is a zero-dimensional temporal region with no temporal parts.

In scientific research, it is relevant to categorize different types of information entities that carry scientific knowledge. For this reason, the Information Artifact Ontology (IAO) is an ontology created using BFO as a basis, capturing numerous information entities existent in scientific research (e.g., protocols, documents, experimental logs, databases, published literature and so on).

Information Content Entity. In IAO, an information content entity is a *generically dependent continuant* that refers to (*is about*) some entity.

Document. A number of information entities that must be understood together as a whole.

As an upper ontology, BFO provides the foundation for over 350 domain ontology extensions in multiple domains [6]. In this paper, BFO is used for augmenting data with semantics. We refrain from including a definition for all the concepts here required in Section 3, referring the reader to [4, 6] for a comprehensive definition.

3. OntoOCEL Framework

This section describes the OntoOCEL framework to augment OCEL event data with process and domain knowledge:

3.1. Step 1: Understand How Process Knowledge Relates to the Ontology

The first step of our methodology consists of conducting a semantic analysis to understand how the process knowledge structure can be mapped to the to upper level ontology, regardless process instances.

In BPM literature, a business process is composed of a set of *activities* executed to handle a case (Section 2.1) (the control-flow perspective). Activities and processes are performed by *roles* that can be fulfilled by people or organization units (resource perspective), manipulating data and information (data perspective), happening during a period of time (time perspective) [1].

An assessment of BFO classes (Section 2.2), together with use cases on how to relate particulars to BFO universals and defined classes defined in [6] lead to the modeling decisions depicted in Figure 1. In short, BPM concepts are mapped to BFO *particulars* (represented as diamonds). These particulars are related to BFO *universals* and *defined classes* (represented as ovals) through an *instance of* relation. The relations among the particulars are defined as follows. Roles (organizational or person) *participate in* activities at a timestamp t. Activities *consume* or *produce* data objects.

Example. The aforementioned modeling decisions are schematically presented in Figure 1.

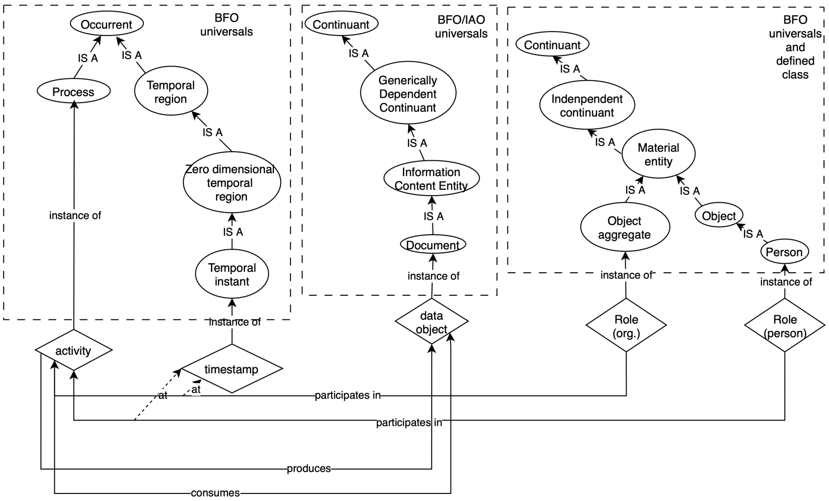


Figure 1: Process Knowledge Mapped to BFO Upper Ontology

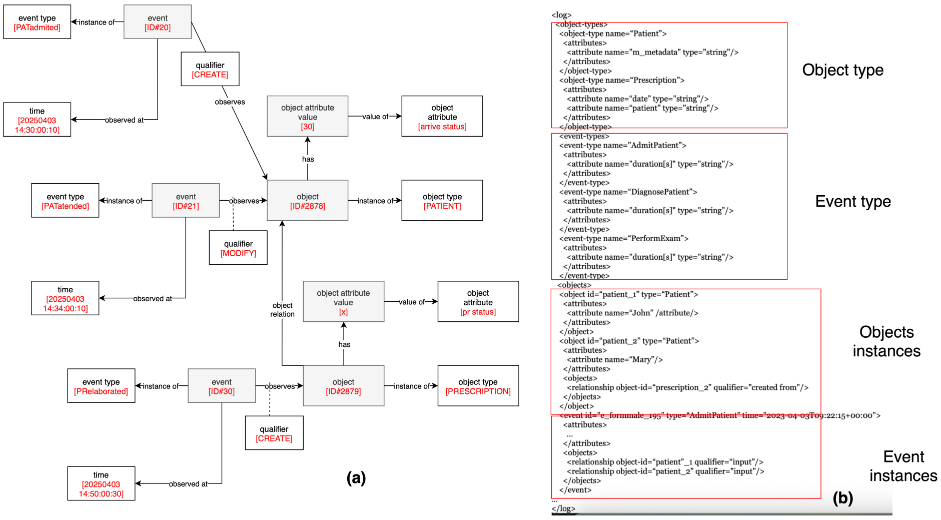
3.2. Step 2: Understand the Semantics of Event Data in OCEL Logs

Now that we know how the process knowledge relates to the upper ontology, we can think about the domain knowledge. To do this, we initially look into the data stored in the event logs.

To facilitate interoperability among process mining tools, event logs are standardized. Figure 2(a) depicts the OCEL metamodel adapted from [3], while Figure 2(b) depicts a sample of synthetic, manually generated OCEL event log. As can be seen in Figure 2, OCEL event logs contain object types (and attributes), object instances, event types (activities) (and attributes) and event type instances.

With the OCEL event logs in hands, one can understand the data objects that we have available data, together with the events that manipulate such data. These data objects will be our domain entities in the next step.

Example. With the event logs in hands, it is possible to identify two data objects specified within the log ($Patient$ and $Prescription$) and one activity ($AdmitPatient$) □



```
<log>
 <object-types>
  <object-type name="Patient">
   <attributes>
    <attribute name="m_metadata" type="string"/>
   </attributes>
  </object-type>
  <object-type name="Prescription">
   <attributes>
    <attribute name="date" type="string"/>
    <attribute name="patient" type="string"/>
   </attributes>
  </object-type>
  <event-types>
  <event-type name="AdmitPatient">
   <attributes>
    <attribute name="duration[s]" type="string"/>
   </attributes>
  </event-type>
  <event-type name="DiagnosePatient">
   <attributes>
    <attribute name="duration[s]" type="string"/>
   </attributes>
  </event-type>
  <event-type name="PerformExam">
   <attributes>
    <attribute name="duration[s]" type="string"/>
   </attributes>
  </event-type>
 <objects>
  <object id="patient_1" type="Patient">
   <attributes>
    <attribute name="John" /attribute/>
   </attributes>
  </object>
  <object id="patient_2" type="Patient">
   <attributes>
    <attribute name="Mary"/>
   </attributes>
   <objects>
    <relationship object-id="prescription_2" qualifier="created from"/>
   </objects>
  </object>
  <event id="e_formmale_195" type="AdmitPatient" time="2023-04-03T09:22:15+00:00">
   <attributes>
    ...
   </attributes>
   <objects>
    <relationship object-id="patient"_1 qualifier="input"/>
    <relationship object-id="patient_2" qualifier="input"/>
   </objects>
  </event>
...
</log>
```

Figure 2: (a) OCEL metamodel adapted from [3] and (b) Sample of synthetic, manually generated OCEL event log

3.3. Step 3: Develop the Domain Ontology Based on Event Data

The third step consists of understanding the nature of knowledge stored in these data objects and how they fit with the knowledge structure of process knowledge (and indirectly to BFO). Activities are instantiated as subclass of BFO *process*, the data objects are instantiated as subclasses of BFO *information content entity*, roles are instantiated either as subclass of (defined class) *person* or as subclass of BFO *object aggregate*. The data objects also carry the values of the data object attributes. These are modeled as OWL Properties.

Example. Figure 3(a) depicts the hierarchy of BFO classes in OWL. First, the OCEL *object types* ($Patient$, $Prescription$ and $Treatment$) are instantiated as subclasses of *information content entity*. As one can see in the hierarchy, BFO provides different types of subclasses as carrier content (e.g. figures, documents, email, etc.). Here, the class $ElectronicRecord$ has been created as subclass of BFO *document* and pointed to class $Patient$ (subclass of BFO *object*), since a BFO information content entity *is about* some entity in reality. Relations among concepts are modeled as OWL Object Properties in Figure 3(b). Activities $AdmitPatient$, $PerformExam$ and $DiagnosePatient$ are created as subclass of BFO *process*. □

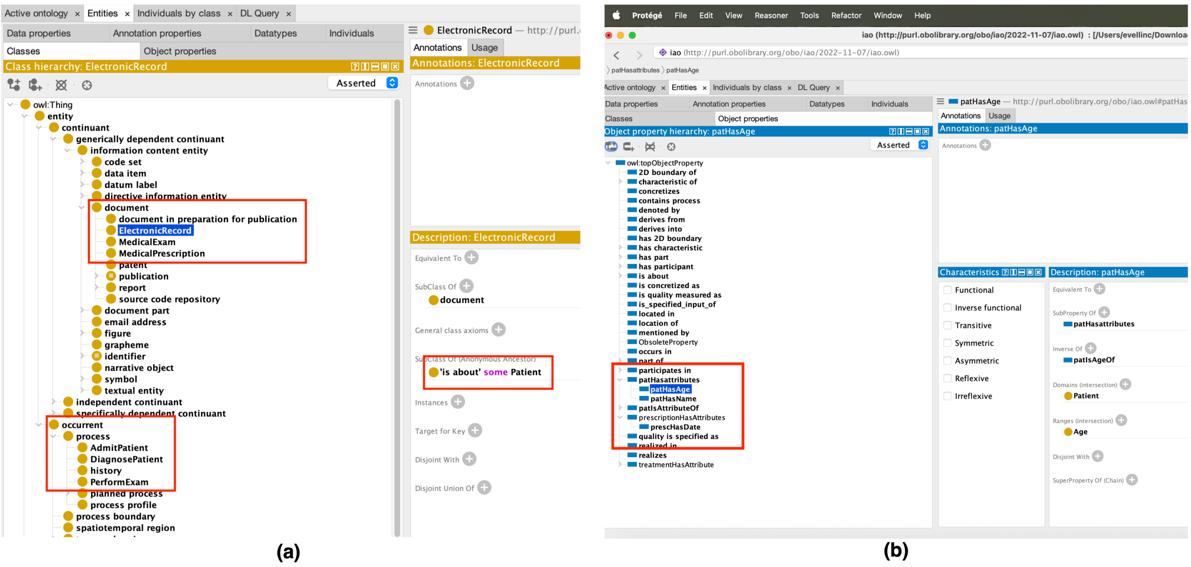


Figure 3: The Domain Ontology Linked to BFO Universals (1)

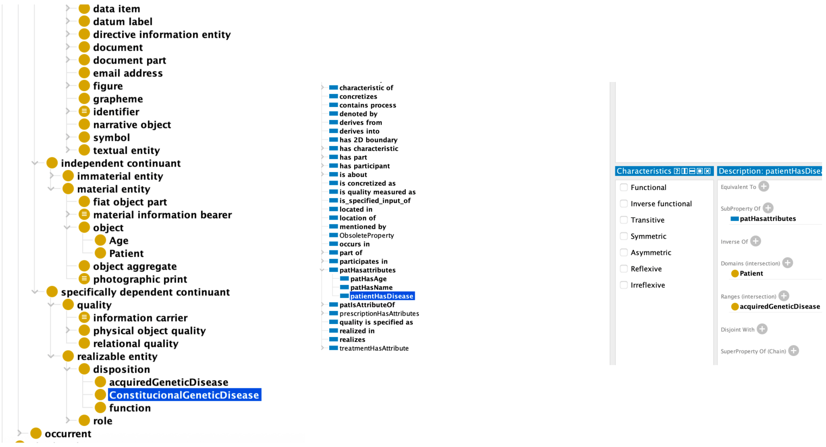
3.4. Step 4: Complement the Domain Ontology with Declarative Knowledge

In this step, the knowledge structure of the enterprise may be complemented with declarative knowledge. This knowledge should be related with the procedural knowledge modeled in the previous steps.

Example. Imagine an admission process that patients may have different types of diseases. The doctor would like to know about it, he wants to track it along time. In this case, we insert a $constitucionalGeneticDisease$ and $acquiredGeneticDisease$ as a disposition as depicted in [4]. Figure 4 depicts this modeling decision. □

3.5. Step 5: Integrate the Data with Ontologies

This step concerns the linkage between processual and domain data with the ontology. The Cellfie plugin [7] may be used for importing the OCEL event logs to the OWL ontology. Working with Cellfie requires the input data to be stored in an Excel spreadsheet, and the creation of *Transformation Rules* that maps the Excel data to the OWL axiom structure. The transformation rules are written in Manchester Syntax. Currently, an event log does not exist for the process, although the ontology has been already conceived in such a way to enable subsequent import of event logs. Further, domain specific data may



□ Functional
□ Inverse functional
□ Transitive
□ Symmetric
□ Asymmetric
□ Reflexive
□ Irreflexive

Equivalent To
SubProperty Of
patHasattributes
Inverse Of
Domains (intersection)
Patient
Ranges (intersection)
acquiredGeneticDisease
Disjoint With
SuperProperty Of (Chain)

Figure 4: The Domain Ontology Linked to BFO Universals (2)

also be inserted into the ontology (e.g. data about the $constitucionalGeneticDisease$), extracted from relational databases, or other sources of information.

4. Related Work

The idea of providing semantics for BPM technologies has been embraced in two veins in literature. A first group of approaches use Ontologies and Semantic Web technologies combined with business process models [8, 9, 10, 11], augmenting the process logic with domain knowledge. These approaches are grounded on domain ontologies and OWL language, which does not embrace a semantic view supported by upper ontologies. The only exception is Pedrinaci et. al. [8] that considers upper ontologies, without deepening further on data.

In a second strain of research that augment event data with Semantic Web technologies [12, 13, 14, 15, 16], Ciccio et. al. [15] and Khayatbashi et. al. [16] are focused on temporal constraints, not on semantics. Eichele et. al. [12] indeed focus on semantics, but only consider domain ontologies (not upper) and classical event logs. Xiong et. al. [13] and Swevels et. al. [14] emphasize integration, extraction, and transformation, with weak semantic foundations and no foundational ontology grounding. Differently, this paper is explicitly dedicated to integrating OCEL data with the Basic Formal Ontology (BFO), proposing a framework that maps OCEL event structures to BFO categories to represent processes, participants, objects, and roles more richly. This approach enhance semantic meaning, interoperability, and support for knowledge-intensive process analysis, since it uses a well-leveraged ontology.

5. Conclusion

This paper has presented a comprehensive framework, OntoOCEL, designed to address the critical semantic gap in Process Mining by augmenting Object-Centric Event Logs (OCEL) with a rich ontological foundation derived from the Basic Formal Ontology (BFO). We have outlined a five-step methodology that facilitates this integration, moving from a semantic analysis of process knowledge in relation to upper-level ontologies to the development of a domain-specific ontology, its complementation with declarative knowledge, and finally, the seamless integration of event data with the constructed ontology.

While the current work lays a conceptual and methodological foundation, it is important to acknowledge that populating the ontology with event data is a limitation of our work that must addressed in future efforts. Future work will focus on the empirical validation of this framework through the actual import and analysis of real-world OCEL event logs, by elaborating the transformation rules, and exploring the potential for automated reasoning and intelligent decision support based on the enriched ontological models. This will ultimately lead to more intelligent, adaptive, and insightful BPMS systems.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-4 in order to: improve writing style. After using this tool, the author reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Aalst, van der, Business Process Management: A Comprehensive Survey, ISRN Software Engineering 2013 (2013).

[2] W. van der Aalst, J. Carmona, Process Mining Handbook, Lecture Notes in Business Information Processing, Springer International Publishing, 2022.

[3] D. Fahland, M. Montali, J. Lebherz, W. M. P. van der Aalst, M. van Asseldonk, P. Blank, L. Bosmans, M. Brenscheidt, C. di Ciccio, A. Delgado, D. Calegari, J. Peeperkorn, E. Verbeek, L. Vugs, M. T. Wynn, Towards a Simple and Extensible Standard for Object-Centric Event Data (OCED) – Core Model, Design Space, and Lessons Learned, 2024. arXiv:2410.14495.

[4] R. Arp, B. Smith, A. D. Spear, Building Ontologies with Basic Formal Ontology, The MIT Press, 2015.

[5] I. Beerepoot, C. Di Ciccio, H. A. Reijers, S. Rinderle-Ma, W. Bandara, Andrea, The Biggest Business Process Management Problems to Solve Before We Die, Computers in Industry 146 (2023) 103837.

[6] J. N. Otte, J. Beverley, A. Ruttenberg, Basic Formal Ontology: Case Studies, Applied Ontology 17 (2022) 17–43. doi:10.3233/ao-220262.

[7] M. O'Connor, Cellfie plugin, https://github.com/protegeproject/cellfie-plugin/wiki/Grocery-Tutorial, 2025.

[8] C. Pedrinaci, J. Domingue, A. K. Alves de Medeiros, A Core Ontology for Business Process Analysis, in: The Semantic Web: Research and Applications, Springer Berlin Heidelberg, 2008, pp. 49–64.

[9] M. Hepp, D. Roman, An Ontology Framework for Semantic Business Process Management, in: Proceedings of Wirtschaftsinformatik, 2007, pp. 423–440.

[10] C. Di Francescomarino, C. Ghidini, M. Rospocher, L. Serafini, P. Tonella, Semantically-Aided Business Process Modeling, in: The Semantic Web - ISWC 2009, Springer Berlin Heidelberg, 2009, pp. 114–129.

[11] P. Bertoli, F. Corcoglioniti, C. Di Francescomarino, M. Dragoni, C. Ghidini, M. Pistore, Semantic modeling and analysis of complex data-aware processes and their executions, Expert Systems with Applications 198 (2022) 116702.

[12] S. Eichele, K. Hinkelmann, M. Spahic-Bogdanovic, Ontology-Driven Enhancement of Process Mining With Domain Knowledge, in: Proceedings of AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), 2023, p. vvv.

[13] J. Xiong, G. Xiao, T. E. Kalyci, M. Montali, Z. Gu, D. Calvanese, A Virtual Knowledge Graph Based Approach for Object-Centric Event Logs Extraction, in: Revised Selected Papers of the Process Mining Workshops (ICPM-WS 2022), volume 468 of *Lecture Notes in Business Information Processing*, Springer, 2022, pp. 466–478. doi:10.1007/978-3-031-27815-0_34.

[14] A. Swevels, D. Fahland, M. Montali, Implementing Object-Centric Event Data Models in Event Knowledge Graphs, in: Process Mining Workshops, Springer Nature, 2024, pp. 431–443.

[15] C. di Ciccio, F. Ekaputra, A. Cecconi, A. Ekelhart, E. Kiesling, Finding Non-Compliances with Declarative Process Constraints Through Semantic Technologies, in: International Conference on Advanced Information Systems Engineering Forum (CAiSE Forum), Springer, 2019, pp. 60 – 74.

[16] S. Khayatbashi, O. Hartig, A. Jalali, Transforming Event Knowledge Graph to Object-Centric Event Logs: A Comparative Study for Multidimensional Process Analysis, in: Proceedings of 42nd International Conference on Conceptual Modeling (ER 2023), Springer-Verlag, Berlin, Heidelberg, 2023, p. 220–238. doi:10.1007/978-3-031-47262-6_12.

TAFNAVEGA as an Interoperable Semantic Vocabulary: From Faceted Taxonomy to Ontological Modeling

Benildes Coura Moreira dos Santos Maculan[1,*,†] and Elisângela Cristina Aganette[2,†]

[1, 2] *Universidade Federal de Minas Gerais, Av. Presid. Antônio Carlos, 6627, Pampulha, BH, MG, CEP 31270-901, Brasil*

Abstract

The navigational faceted taxonomy TAFNAVEGA was originally proposed as a categorical structure for organizing theses and dissertations. With the advancement of Open Science and the Semantic Web, it has become necessary to revisit this proposal to align it with contemporary standards of formal representation and semantic interoperability. This study, which is qualitative, exploratory, and conceptual in nature, presents the reformulation of TAFNAVEGA based on Faceted Classification Theory, semantic standards such as SKOS (Simple Knowledge Organization System) and OWL (Web Ontology Language), and in dialogue with the ontological model CERIF (Common European Research Information Format). The methodology comprised a diagnosis of the original structure (CAFTE), the mapping of semantic relations and logical dependencies, and the alignment with formal representation standards. The results, still at a conceptual stage, indicate the feasibility of structuring CAFTE as an interoperable controlled vocabulary, while highlighting the need for conceptual refinements, computational formalization, and empirical validation in thesis and dissertation repositories. The proposal underscores TAFNAVEGA's contribution to academic knowledge organization and its potential integration into knowledge graphs aimed at Open Science and Artificial Intelligence.

Keywords

Faceted taxonomy; Knowledge organization; Semantic interoperability; Academic ontologies.

1. Introduction

The growing adoption of digital libraries by academic and research institutions has intensified the demand for classificatory structures that enable more precise and semantically oriented information retrieval, consistent with the logic of scientific production. This challenge is particularly evident in repositories of theses and dissertations, given the diversity of topics, methods, and objects of investigation.

In this context, the navigational faceted taxonomy (TAFNAVEGA), proposed by Maculan [1], emerged as a pioneering initiative of categorical organization, structured around ten facets. Its formulation was based on Bardin's thematic categorical analysis [2] and inspired by Ranganathan's Faceted Classification Theory (FCT) [3], providing a strategy for navigable representation of academic content.

More than a decade later, the principles of Open Science, the consolidation of the FAIR standards —Findable, Accessible, Interoperable, and Reusable [4]—and the growing adoption of semantic standards such as SKOS (Simple Knowledge Organization System) [5] and OWL (Web Ontology Language) [6] highlight the need to revisit TAFNAVEGA. In this scenario, models oriented to the formal representation of scientific activity gain relevance, among which CERIF (Common European Research Information Format) [7] stands out, widely used in European Current Research Information Systems (CRIS).

Based on this framework, the general objective of this article is to propose the reformulation of TAFNAVEGA as an interoperable controlled vocabulary, grounded in FCT and formal semantic

[1] *Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.*

[*] Corresponding author.

[†] These authors contributed equally.

benildes@gmail.com (B.C.M.S. Maculan); elisangelaaganette@gmail.com (E. C. Aganette).

0000-0003-4303-9071 (B.C.M.S. Maculan); 0000-0003-4357-8016 (E. C. Aganette).

standards. Specifically, the study seeks to: (i) diagnose the limitations of the original structure; (ii) map its facets in dialogue with the CERIF ontological model; (iii) propose guidelines for formalization in SKOS and OWL; and (iv) outline perspectives for application in academic repositories. The research question guiding the study is: How can TAFNAVEGA be updated to operate as a semantically interoperable component in Open Science ecosystems?

2. Theoretical and methodological foundations

The construction of interoperable classificatory systems in digital environments requires a theoretical-methodological basis that combines consolidated principles of Knowledge Organization (KO) with approaches to formal representation. The Faceted Classification Theory (FCT), originating in Librarianship and consolidated in Information Science, remains a relevant analytical-synthetic model in the face of challenges posed by the Semantic Web and Artificial Intelligence. Its ability to decompose and recombine concepts into multidimensional facets makes it a reference for the development of faceted taxonomies and lightweight ontologies, applicable in interdisciplinary contexts.

This section presents four key axes that underpin the reformulation of TAFNAVEGA: (i) a critical review of FCT (2.1); (ii) description of the categorical matrix CAFTE, derived from empirical analysis of theses and dissertations (2.2); (iii) presentation of the original TAFNAVEGA proposal and its potential for reformulation (2.3); and (iv) discussion of SKOS and OWL standards as technical references for its formalization (2.4).

By articulating these foundations, the study emphasizes the dialogue between KO traditions and Semantic Web models, providing the conceptual basis for the reformulation of TAFNAVEGA and its prospective application in Artificial Intelligence and knowledge representation systems.

2.1. Classification Theory and its application in Information Science

Proposed by Ranganathan in the 1930s, Faceted Classification Theory (FCT) is one of the most influential foundations of Knowledge Organization (KO). By decomposing subjects into five fundamental categories — Personality, Matter, Energy, Space, and Time (PMEST) — it broke with traditional enumerative models, offering a flexible and synthetic structure. Despite its conceptual robustness, FCT lacks formal computational mechanisms, which highlights the importance of integrating it with Semantic Web standards such as SKOS and OWL.

Recent studies confirm its relevance. Silva and Miranda [8] emphasize its precision and adaptability for representing interdisciplinary domains, while Coelho, Lima and Borges [9] point to the effectiveness of faceted taxonomies for semantic retrieval in digital environments, particularly when formalized in interoperable standards. Almeida and Teixeira [10] argue that, in educational contexts, faceted classification also functions as an epistemic instrument, enhancing usability and reusability of digital objects. Lima [11] highlights its compatibility with the epistemological diversity of scientific knowledge, reinforcing its pertinence for thesauri and interoperable taxonomies.

Some proposals have sought to update the theory, such as Martins, Moreira and Santarém Segundo [12], who suggest alternative fundamental categories (thing, cultural object, event, person, place, time, attribute, action, cause, and purpose), thus incorporating dimensions such as causality and intentionality. This effort resonates with the reformulation of TAFNAVEGA, reinforcing the potential of aligning faceted structures with ontological models and semantic standards, thereby expanding their interoperability and applicability in scientific information systems.

2.2. The categorical matrix: construction process

The categorical matrix, named CAFTE (Faceted Analytical Classification of Theses and Dissertations), was developed by Maculan [1] to represent, with granularity and flexibility, the research objects found in theses and dissertations. Its construction was theoretically based on FCT, especially Ranganathan's PMEST model [3], and methodologically on qualitative categorical content analysis [2].

The corpus consisted of titles, abstracts, and keywords from 41 documents in the UFMG Institutional Repository, within the Information Organization research line. The process involved: (i) corpus collection and selection; (ii) coding into emerging thematic categories; (iii) grouping by semantic proximity and discursive function; (iv) definition of facets and subfacets; and (v) experimental validation [1]. This ensured that categories were empirically derived from actual scientific discourse, aligning representation with the internal logic of the texts.

The outcome was the definition of ten functional categories: Theme, Empirical object, Scope, Context, Research type, Data collection, Methods, Theoretical framework, Historical/contextual framework, and Results [1]. Although not identical to PMEST, correspondences with Ranganathan's proposal were established, reaffirming CAFTE's affinity with the faceted tradition of KO.

Originally conceived to support indexing and retrieval in academic repositories, CAFTE also demonstrates potential for more complex computational environments. Its faceted structure can be formalized in SKOS or OWL, becoming an ontologically grounded controlled vocabulary capable of semantic interoperability. As such, it serves as a bridge between empirical conceptual extraction and lightweight ontological modeling, supporting applications in knowledge graphs, recommender systems, and generative AI.

2.3. The original TAFNAVEGA proposal: structure, categories, and implementation

TAFNAVEGA was conceived from the CAFTE matrix [1] as a mechanism for representing and accessing theses and dissertations, aiming to enable faceted navigational exploration rather than simple keyword search, often constrained by ambiguity and low precision.

Its initial implementation took place through a prototype developed in Microsoft Access [16], designed according to the principles of multidimensional classification and faceted navigation. Each CAFTE facet generated a set of terms organized into independent tables, allowing progressive combinations. The interface enabled users to refine queries by selecting multiple categories, such as identifying qualitative research using interviews in the health field, grounded in Activity Theory.

Although it did not incorporate formal representation languages such as RDF, SKOS, or OWL, TAFNAVEGA demonstrated the feasibility of faceted navigation in academic repositories. However, the absence of notations, URIs, and explicit ontological relations limited its interoperability. The reformulation proposed in this study seeks to advance its formalization in SKOS and OWL, expanding its applicability to institutional repositories and broader Open Science ecosystems.

2.4. SKOS and OWL semantic patterns

SKOS and OWL, developed by the W3C, are key standards for representing structured knowledge and enabling semantic interoperability on the Web. In the field of Knowledge Organization, they have been widely adopted for modeling controlled vocabularies, classification schemes, and ontologies.

SKOS provides a lightweight framework for representing concepts and basic semantic relations, such as hierarchy (skos:broader, skos:narrower) and association (skos:related). Its simplicity and machine readability make it especially suitable for expressing taxonomies, thesauri, and faceted schemes in an interoperable way [5].

OWL, in contrast, is a more expressive language grounded in description logic, capable of defining classes, properties, equivalence or disjointness relations, and formal restrictions [6]. These features make it particularly useful in contexts that require automatic inference, consistency checking, and integration with intelligent agents.

In this study, SKOS is considered appropriate for structuring TAFNAVEGA's categories and basic relations, while OWL is proposed as an additional layer to formalize logical dependencies among facets. For instance, the facet C5: Research Type can be represented in SKOS as a skos:Concept, with terms linked through hierarchical relations, while constraints between complementary facets can be expressed in OWL through axioms.

The combined use of SKOS and OWL thus offers a balanced approach between simplicity and expressiveness, enabling the reformulation of TAFNAVEGA as an interoperable controlled vocabulary aligned with the Semantic Web.

3. Methodology for the Reformulation of TAFNAVEGA

The reformulation of TAFNAVEGA followed a qualitative, exploratory, and descriptive approach, based on conceptual modeling and manual semantic analysis. No automated tools were employed; instead, the work relied on the researchers' experience with faceted taxonomies, theoretical frameworks in Knowledge Organization and the Semantic Web, and a critical review of relevant literature. Conceptual decisions were documented in analytical matrices and iteratively revised to ensure coherence and minimize bias.

The proposal builds on the original TAFNAVEGA, developed by Maculan [1] through empirical analysis of theses and dissertations, and focuses on its conceptual and technical update. Rather than collecting new data, we conducted semantic comparisons of categorical structures. Facets were examined against PMEST and CERIF models to identify terminological alignments, logical dependencies, and opportunities for representation in SKOS and OWL. The methodological process comprised four stages:

Stage 1 – Diagnosis of the existing structure: critical analysis of CAFTE and the first version of TAFNAVEGA, assessing semantic clarity, coverage, and relevance for representing academic objects. Comparative analyses were made with PMEST and CERIF [7]. CERIF was selected as a reference due to its recognition as an international standard promoted by euroCRIS and its RDF/OWL compatibility [6], though it was not fully adopted to avoid scope limitations.

Stage 2 – Identification of semantic relations: hierarchical, associative, and equivalence relations among categories were mapped, together with logical dependencies across facets, to support coherent navigation and retrieval.

Stage 3 – Mapping to formal standards: SKOS schemes were designed to model facets as ConceptSchemes and terms as Concepts, with semantic properties explicitly defined. OWL was also considered for representing more complex dependencies relevant to automated inference.

Stage 4 – Alignment with external ontologies: equivalences between CAFTE categories and CERIF entities [7] were established to strengthen interoperability in Open Science ecosystems. This alignment supports applications such as content recommendation, semantic enrichment of metadata, and generative AI in academic repositories. ChatGPT 4.0 [13] was used as a heuristic aid for semantic exploration, with results validated by the authors.

As a conceptual exercise, this proposal does not include empirical validation with users. Future work will address this gap through pilot studies in institutional repositories and retrieval experiments. Nonetheless, the methodology provides a consistent conceptual basis—anchored in Information Science and Semantic Web frameworks—for reformulating TAFNAVEGA as an interoperable controlled vocabulary for Open Science environments.

4. Results

The reformulation of TAFNAVEGA focuses on creating a more refined semantic artifact, oriented towards interoperability and advanced computational use. Results are organized according to the four methodological steps.

4.1. Diagnosis and validation of the existing categorical structure

The critical analysis of CAFTE, the matrix that originated TAFNAVEGA, was conducted through a comparison between its ten categories, the classical facets of Ranganathan's Faceted Classification Theory (PMEST) [3], and the components of the CERIF model (Common European Research Information Format) [7]. While Maculan [1, p. 142, table 13] had already established parity between CAFTE and PMEST, this study adds the novel alignment with CERIF.

Correspondences between CAFTE and CERIF were established by conceptual inference, using three criteria: functional equivalence (same structural role with different terms), semantic equivalence (similar meaning with different functions), and mixed equivalence (function + meaning). Table 1 summarizes these mappings.

Table 1
Comparative alignment CAFTE (TAFNAVEGA) × PMEST × CERIF

CAFTE (TAFNAVEGA)	Ranganathan (PMEST)	CERIF (Common European Research Information Format)
1. Theme	Personality	cfResPubl_Class (thematic classification)
2. Object	Personality	cfResPubl / cfResProd / cfResPat (research product/object)
3. Results	Personality	cfResultProduct (scientific output)
4. Scope	Matter	cfResPubl / cfProj (Theme, scope, domain)
5. Historical/contextual foundation	Matter	cfProj / cfEvent (historical context, project, event)
6. Research type	Matter	cfProj / cfResPublClassification
7. Data collection	Energy	cfResPubl / cfMethod (applied methods)
8. Theoretical foundation	Energy	cfTheory / cfResPubl (conceptual base)
9. Methods	Energy	cfMethod / cfTechnique (applied technique)
10. Setting	Espace	cfOrgUnit / cfFacility / cfPlace (institutional or spatial context)

Source: Authors (2025).

Table 2 shows that hierarchical relationships were represented by generalization/specialization structures in SKOS, while associative connections and functional dependencies required greater expressiveness in OWL. This conceptual modeling was defined manually, by inference of the authors, which constitutes a theoretical exercise not yet validated on empirical bases.

In line with CERIF, it was observed that although the model is not classificatory, its relational architecture allows for the incorporation of extensions and external vocabularies. In this sense, the semantic links identified in TAFNAVEGA can be associated with entities such as cfResPubl, cfProj, cfResultProduct, and cfMethod, expanding the potential for semantic interoperability with CRIS systems and knowledge graphs. It should be noted, however, that this alignment is exploratory: it serves as a basis for integration, but does not replace the need for formalization in SKOS/OWL.

The explicitation of semantic relationships strengthens the conceptual consistency of TAFNAVEGA and guides its future computational implementation. In addition to technical robustness, it is recommended that the next steps include interface prototypes in academic repositories in order to empirically test the effectiveness of faceted navigation, as well as the added value to the semantic retrieval of scientific information. Table 2 shows that hierarchical relationships were represented by generalization/specialization structures in SKOS, while associative connections and functional dependencies required greater expressiveness in OWL. This conceptual modeling was defined manually, by inference of the authors, which constitutes a theoretical exercise not yet validated on empirical bases.

In alignment with CERIF, it was observed that although the model is not classificatory, its relational architecture allows for the incorporation of extensions and external vocabularies. In this sense, the semantic links identified in TAFNAVEGA can be associated with entities such as cfResPubl, cfProj, cfResultProduct, and cfMethod, expanding the potential for semantic interoperability with CRIS systems and knowledge graphs. It should be noted, however, that this alignment is exploratory: it serves as a basis for integration, but does not replace the need for formalization in SKOS/OWL.

The explicitation of semantic relationships strengthens the conceptual consistency of TAFNAVEGA and guides its future computational implementation. In addition to technical robustness, it is recommended that the next steps include interface prototypes in academic repositories in order to empirically test the effectiveness of faceted navigation, as well as the added value to the semantic retrieval of scientific information.

4.1.1. Conceptual Convergences

CAFTE categories show strong alignment with TCF principles, particularly the PMEST structure. Facets such as Theme, Object, and Results correspond to the Personality category, central to subject determination in PMEST. Similarly, Methods, Theoretical foundation, and Data collection align with the Energy facet, representing processes applied to research objects.

In the CERIF model, widely used for standardized descriptions of scientific activity, relevant correspondences also emerge:

- Theme ↔ cfResPubl_Class (thematic classification of scientific output);
- Object ↔ cfResPubl / cfResProd / cfResPat (research products or objects);
- Results ↔ cfResultProduct (formal representation of research outcomes);
- Methods and Data collection ↔ cfMethod and cfTechnique;
- Context ↔ cfOrgUnit, cfFacility, or cfPlace.

These mappings support the feasibility of TAFNAVEGA as an interoperable vocabulary, capable of dialoguing with established standards in scientific information. Full adoption of a European model such as CERIF, however, may require contextual adjustments for specific domains, particularly in Brazil, given institutional, cultural, and linguistic differences. Thus, the mapping should be seen as an exploratory reference, open to adaptation and extension.

4.1.2. Gaps and Misalignments

The comparative analysis also revealed some conceptual misalignments. The Results facet, for instance, originally tied to Personality, is better understood as a dimension of Energy, as it denotes the outcome of a methodological process. In CERIF, although cfResultProduct reflects this aspect, it does not capture the causal relationship inherent in the PMEST model.

Other limitations concern more abstract elements, such as Theoretical Foundation and Historical Context, which in CERIF appear only as descriptive fields (e.g., cfResPubl, cfProj), without dedicated entities or explicit hierarchies. Likewise, the Setting facet, central to CAFTE, is fragmented across multiple entities (cfOrgUnit, cfPlace, cfFacility), hindering an integrated and cohesive representation.

These limits show that CAFTE–CERIF compatibility, while strong, is not complete. To bridge these gaps, we recommend: (a) pilot tests in institutional repositories, applying the mapping to real dissertation records to validate accuracy and resolve ambiguities; and (b) complementary ontological extensions in SKOS and OWL, especially for theoretical foundations and historical contexts, which require richer semantic expressivity.

Overall, the analysis confirms the potential for convergence but also highlights the need to complement CERIF with richer formal ontologies to ensure both interoperability and the conceptual expressiveness of TAFNAVEGA.

4.2. Semantic Relations and Logical Dependencies

In the second stage, we developed a model of semantic relationships between the categories of the reformulated TAFNAVEGA, using the standard properties of controlled vocabularies as a reference (broader, narrower, related, and equivalent). The process was documented in categorical comparison spreadsheets, which contained descriptions, examples, and critical comments to ensure transparency and auditability for future external validations. The analysis resulted in the systematization of hierarchical, associative, and functional relationships, which are presented in Table 2 along with implications for their formalization in SKOS and OWL.

Table 2
Semantic relations and logical dependencies of CAFTE categories

Category (CAFTE)	**Type of relation**	**Example of Dependency or Association**	**Implication for SKOS / OWL**
Theme	Hierarchical	'Education' > 'Inclusive education'	skos:broader/narrower
Object	Associative	''Teachers' related to 'Teacher training'	skos:related; can be owl:ObjectProperty in complex graphs
Results	Functional dependency with Methods	'Predictive model' generated by 'Statistical analysis'	owl:hasOutput; owl:Restriction
Scope	Hierarchical / Associative	'School libraries' narrower than 'Libraries'	skos:broader; skos:related for adjacent scopes
Historical/contextual foundation	Associative	'Student movement' related to 'Educational reforms'	skos:related; can be integrated with dc:coverage (temporal)
Type of research	Hierarchical	'Case study' narrower than 'Qualitative research'	skos:broader/narrower
Data collection	Dependency with Research type	'Interviews' requires 'field research'	owl:Restriction; conditional logic with owl:someValuesFrom
Theoretical foundation	Associative	'Freire' related to 'Critical pedagogy'	skos:related; owl:equivalentClass between theoretical schools
Methods	Hierarchical / dependency	'Content analysis' narrower than 'Qualitative analysis'	skos:broader; relationship with data type or support via OWL
Setting	Dependency with Population and Theme	'Public school' where 'pedagogical practice' takes place	owl:hasLocation; or skos:related to Entity and Theme

Source: Authors (2025).

Table 2 shows that hierarchical relationships are represented by generalization/specialization structures in SKOS, while associative links and functional dependencies require the greater expressiveness of OWL. This modeling was defined manually by the authors, as a theoretical exercise not yet empirically validated.

In line with CERIF, although not a classificatory model, its relational architecture supports extensions and external vocabularies. Accordingly, the semantic links identified in TAFNAVEGA can be mapped to entities such as cfResPubl, cfProj, cfResultProduct, and cfMethod, enhancing semantic interoperability with CRIS systems and knowledge graphs. This alignment, however, remains exploratory and does not replace the need for formalization in SKOS/OWL.

Explicit semantic relationships strengthen TAFNAVEGA's conceptual consistency and guide future computational implementation. Next steps should include interface prototypes in academic repositories to empirically test the effectiveness of faceted navigation and its added value for semantic retrieval of scientific information.

4.3. Mapping to Conceptual Representation Standards

Based on the semantic relations and logical dependencies identified in the previous stage, an initial conceptual modeling of TAFNAVEGA was developed according to SKOS principles. This modeling remains at an exploratory level, without implementation in RDF language, and aims to project the future formalization of TAFNAVEGA as an interoperable controlled vocabulary.

Each facet was conceived as a potential skos:ConceptScheme, while the terms were treated as skos:Concept. At the theoretical level, preferred and alternative labels were defined (skos:prefLabel, skos:altLabel), as well as definitions (skos:definition) and hierarchical or associative relations (skos:broader, skos:narrower, skos:related). Figure 1 illustrates an example of modeling for the CAFTE Research Type category [13].

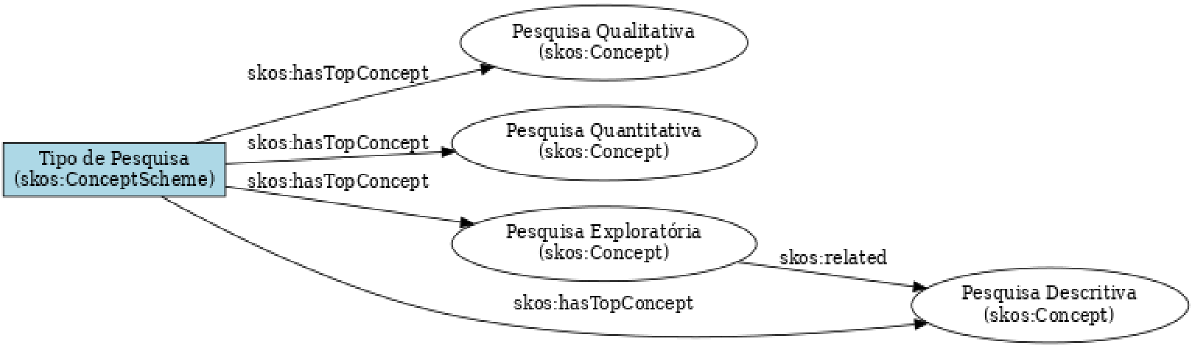

Tipo de Pesquisa (skos:ConceptScheme)
skos:hasTopConcept
Pesquisa Qualitativa (skos:Concept)
skos:hasTopConcept
Pesquisa Quantitativa (skos:Concept)
skos:hasTopConcept
Pesquisa Exploratória (skos:Concept)
skos:related
skos:hasTopConcept
Pesquisa Descritiva (skos:Concept)

Figure 1 – Example of SKOS modeling for the CAFTE Research Type category [13].

In addition to SKOS, the complementary application of OWL was considered, particularly to express restrictions, functional properties, and logical inferences across facets. Table 3 summarizes the main mapping implications for each CAFTE category.

Table 3
Possibilities for formal representation of CAFTE categories in SKOS and OWL

CAFTE Category	Implication for SKOS / OWL
Theme	skos:broader/narrower
Object	skos:related; may be owl:ObjectProperty in complex graphs
Results	owl:hasOutput; owl:Restriction
Scope	skos:broader; skos:related for adjacent scopes
Historical/contextual foundation	skos:related; may integrate with dc:coverage (temporal)
Research type	skos:broader/narrower
Data collection	owl:Restriction; conditional logic with owl:someValuesFrom
Theoretical foundation	skos:related; owl:equivalentClass between theoretical schools
Methods	skos:broader; related to data type or medium via OWL
Setting	owl:hasLocation; or skos:related to Entity and Theme

Source: the authors (2025).

The choice between SKOS and OWL was guided by the nature of the relations to be represented. Hierarchical, associative, or equivalence relations were modeled in SKOS, given its simplicity and broad interoperability. In turn, functional dependencies, co-occurrence conditions, and logical restrictions (such as "requires," "generates," or "occurs in") were projected in OWL, ensuring semantic expressiveness to support automatic inference and integration with AI systems.

Although still at a conceptual stage, the results of this phase indicate promising paths for the computational implementation of TAFNAVEGA in environments compatible with the Semantic Web and aligned with FAIR principles, reconciling representational simplicity with semantic robustness.

4.4. Alignment with external ontologies and knowledge graphs

Based on the conceptual structure refined in the previous stages, we outlined possibilities for aligning TAFNAVEGA with the CERIF (Common European Research Information Format) model, a widely adopted standard for representing scientific and academic research activities. Although not originally faceted, CERIF has a relational and modular architecture that supports the incorporation of external vocabularies through RDF/OWL extensions, enabling integration with TAFNAVEGA's categorial logic.

In this mapping exercise, CAFTE categories were associated with core CERIF entities, such as:

- **cfResPubl** – scientific publications, related to Theme, Object, and Results;
- **cfProj** – research projects, aligned with Scope, Research Type, and Theoretical Foundation;
- **cfResultProduct** – research outputs, linked to Results and Methods;
- **cfMethod** – applied methodologies, associated with Data Collection and Procedures;
- as well as other entities addressing infrastructure, historical context, and research agents.

The semantic and logical relations defined in previous stages—hierarchical, associative, and dependency—support this conceptual alignment, projecting TAFNAVEGA as an auxiliary semantic layer in CERIF-based systems. This integration may contribute to: (i) semantic enrichment of metadata for publications and projects; (ii) interoperability across institutional repositories; and (iii) support for recommendation, automated classification, and thematic summarization in open science and AI-driven environments.

Table 4
Correspondence between CERIF entities and CAFTE categories (TAFNAVEGA)

CERIF Entity	CAFTE Category	Notes
cfResPubl	Results	Research outputs (papers, theses, reports)
cfProj	Theme / Object / Scope	Research project as thematic and organizational context
cfResultProduct	Results	Tangible and intangible outputs (models, software, theories)
cfMethod	Methods / Data collection	Scientific methods and applied instruments
cfFund	Theoretical foundation / Historical-contextual foundation	Conceptual and contextual bases of research
cfFacility	Setting	Infrastructure and locations involved in research
cfOrgUnit	Setting / Object	Organizational unit where research takes place
cfPers	Object	Researchers as active agents
cfClassScheme	Research type	Classification scheme representing methodological approaches

Source: the authors (2025).

CERIF, like other widely used standards (e.g., UNESCO Thesaurus, Frascati Manual Taxonomy), supports integration through RDF/OWL mappings. In this scenario, TAFNAVEGA positions itself as a complementary thematic vocabulary, grounded in a multidimensional categorial base, suitable for supporting semantic indexing and retrieval in thesis and dissertation repositories.

As a next step, we propose a pilot study in a real CERIF environment, in partnership with universities or repository consortia. This test will assess the robustness of the alignment, identify potential conceptual adjustments, and validate TAFNAVEGA's applicability in line with open science requirements and FAIR principles.

5. Conclusions

The reformulation of TAFNAVEGA presented in this article constitutes a conceptual sketch, developed through theoretical modeling and intellectual inference. Although the results point to promising directions for the semantic structuring of academic repositories, the current stage remains abstract and lacks practical validation. The semantic relationships between categories, the mappings with the CERIF model, and the preliminary modeling in SKOS/OWL represent a starting point rather than a finalized implementation.

It is acknowledged that the absence of empirical experimentation limits the generalization of the findings. Challenges such as terminological variability across institutions, metadata curation, and the costs of large-scale implementation constitute limitations that should be addressed in future work.

Among the perspectives for continuity, the following stand out: (a) conceptual adjustments in the facets and terms of the taxonomy; (b) application of the reformulated TAFNAVEGA in real subsets of theses and dissertations, aiming to test its effectiveness in information retrieval; (c) integration of the taxonomy into experimental repositories, assessing its performance in categorization, recommendation, and semantic enrichment systems; (d) usability studies with librarians, managers, and end users, to ensure that the structure is intuitive, flexible, and aligned with real needs.

From a theoretical standpoint, this study dialogues with the proposal of fundamental categories presented by Pereira, Moreira, and Santarém Segundo [12], whose emphasis on dimensions such as cause and purpose broadens the repertoire of faceted classification. This approach demonstrates that CAFTE can be refined and extended, consolidating itself as a structure capable of interoperability with contemporary ontological models.

The original contribution of this study thus lies in the articulation between classical faceted taxonomies and interoperable semantic standards, suggesting a hybrid path between classifications and lightweight ontologies applicable to Information Science. Such an approach expands the possibilities of integration with digital systems and strengthens the agenda of Open Science by promoting the standardized semantic description of data and publications.

It is concluded that aligning faceted taxonomies with Semantic Web standards requires continuous and contextualized revisions, but initiatives such as TAFNAVEGA can decisively contribute to interoperability policies, metadata enrichment, and the consolidation of knowledge graphs applied to scientific information.

Acknowledgments

The authors acknowledge the support of the National Council for Scientific and Technological Development (CNPq), Brazil, for funding this research (Processes Author 1 and Author 2).

Declaration on Generative AI

In preparing this work, the authors used ChatGPT-4 as a complementary support tool for the preliminary identification of semantic relations and correspondences between classificatory models, within the context of the conceptual reformulation of TAFNAVEGA. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final version of this publication.

References

[1] B. C. M. S. Maculan, *Taxonomia facetada navegacional: construção a partir de uma matriz categorial para trabalhos acadêmicos*, M.S. thesis, Programa de Pós-Graduação em Ciência da Informação, Univ. Fed. Minas Gerais, Belo Horizonte, Brazil, 2011.

[2] L. Bardin, *Análise de conteúdo*, 4th ed. Lisboa: Edições 70, 2009.

[3] S. R. Ranganathan, *Prolegomena to Library Classification*, 3rd ed. London: Asia Publishing House, 1967. [Online]. Available: http://dlist.sir.arizona.edu/arizona/handle/10150/106370. [Accessed: Jul. 8, 2025].

[4] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016. DOI: 10.1038/sdata.2016.18.

[5] A. Miles and D. Beckett, "SKOS Simple Knowledge Organization System Reference," W3C Recommendation, 2009. [Online]. Available: https://www.w3.org/TR/skos-reference/. [Accessed: Jul. 10, 2025].

[6] M. K. Smith, C. Welty, and D. L. McGuinness, "OWL Web Ontology Language Guide," W3C Recommendation, 2004. [Online]. Available: https://www.w3.org/TR/owl-guide/. [Accessed: Jul. 10, 2025].

[7] euroCRIS, "Common European Research Information Format (CERIF)," 2021. [Online]. Available: https://www.eurocris.org/cerif. [Accessed: Jul. 10, 2025].

[8] M. B. da Silva and Z. D. de Miranda, "Estudo teórico-analítico-sintético sobre a presença de facetas na organização da informação: do físico ao digital," in *Proc. 19th ENANCIB*, Londrina, Brazil, 2018. [Online]. Available: https://proceedings.ci.inf.br/index.php/enancib/article/view/1621. [Accessed: Jul. 8, 2025].

[9] A. G. Coelho, G. Â. de Lima, and M. M. Borges, "As taxonomias navegacionais facetadas e a produção científica da Ciência da Informação: tendências temática e diacrónica (2011–2020)," in *Proc. 6th Encontro Nacional da ISKO Portugal*, Coimbra, Portugal, 2021, pp. 617–633.

[10] M. B. Almeida and L. M. D. Teixeira, "Revisitando os fundamentos da classificação: uma análise crítica sobre teorias do passado e do presente," *Perspect. Ciênc. Inf.*, vol. 25, no. esp., pp. 28–56, Feb. 2020.

[11] G. Â. de Lima, "Organização e representação do conhecimento e da informação na web: teorias e técnicas," *Perspect. Ciênc. Inf.*, vol. 25, no. esp., pp. 57–97, Feb. 2020.

[12] C. M. Pereira, W. Moreira, and J. E. Santarem Segundo, "Classificação facetada: proposta de categorias fundamentais para organizar teses e dissertações em uma biblioteca digital," *Encontros Bibli*, vol. 26, e79427, 2021. DOI: 10.5007/1518-2924.2021.e79427.

[13] OpenAI, *ChatGPT* (versão 4.0), San Francisco, CA, USA: OpenAI, 2023. [Online]. Available: https://chat.openai.com/

[14] N. Oddone and M. Y. F. S. de F. Gomes, "Uma nova taxonomia para a ciência da informação," in *Proc. 5th Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB)*, Belo Horizonte, Brazil, 2003, pp. 1–24.

[15] G. M. Sacco and Y. Tzitzikas, Eds., *Dynamic Taxonomies and Faceted Search: Theory, Practice and Experience.* Berlin, Germany: Springer, 2009.

Integration of the OnDBTuning Ontology into the OuterTuning Framework

Eric Ruas Leão[1], Edward Hermann Haeusler[1] and Sergio Lifschitz[1,†]

[1] *Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro/RJ, Brazil*

Abstract

This paper presents the integration of the *OnDBTuning* ontology into the *OuterTuning* framework to deliver an explainable, semi-automatic approach to relational database tuning. OnDBTuning encodes expert knowledge about indexes, materialised views, and other optimisation actions as SPARQL rules, while OuterTuning captures workloads, runs what-if simulations, and provides an interactive GUI for DBAs. By linking both tools through a RESTful API that converts workload metadata to RDF and returns inferred actions, we replace OuterTuning's rigid heuristics with ontology-driven reasoning. Proprietary SPIN functions were reimplemented in standard SPARQL, removing license restrictions and increasing portability. Experiments with the TPC-H benchmark (scale factor 0.01) show that a curated subset of ontology-suggested indexes improves the Queries-per-Hour metric by 295% over the no-index baseline, while using only 25% of the storage consumed by a reinforcement-learning approach. The resulting architecture offers transparency, modularity, and tangible performance gains, making it suitable for production environments and database-tuning education.

Keywords

Ontology, OnDBTuning, OuterTuning, Database Tuning, SPARQL, Semantic Web

1. Introduction

Database tuning is crucial for optimizing the performance of relational database systems, directly influencing query response times and transaction efficiency [1]. Traditionally, database administrators (DBAs) rely on their expertise and automatic tools to analyze the system's behavior and propose adjustments such as index creation, memory parameter configurations, and query optimizations. However, these methods often involve subjective decisions, lack transparency, and can be inconsistent or limited in adaptability.

To overcome these limitations, ontologies have emerged as a promising approach for fine-tuning databases by explicitly representing domain knowledge and enabling automatic inference through predefined rules [2][3]. The OnDBTuning ontology, designed explicitly for relational database tuning, formalizes and structures this knowledge, enabling adaptive suggestions for various database scenarios[4]. Concurrently, OuterTuning provides DBAs with a modular framework capable of real-time workload monitoring and transparent decision support through visual recommendations [5].

However, OuterTuning initially employed hard-coded heuristics, limiting flexibility and adaptability. Integrating OnDBTuning into OuterTuning addresses this issue by providing dynamic and adjustable inferences, enhancing both tools' capabilities. This integration required reimplementing proprietary SPIN functions in standard SPARQL, which removed dependencies on licensed components and increased portability. The result is a comprehensive, automated solution capable of transparent and justifiable tuning actions, suitable for both professional and educational contexts. In this sense, this work aims to integrate these two previous studies to create a robust tool combining the strengths of OnDBTuning and OuterTuning.

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

[†] These authors contributed equally.

eric.leao@inf.puc-rio.br (E. R. Leão); hermann@inf.puc-rio.br (E. H. Haeusler); sergio@inf.puc-rio.br (S. Lifschitz)
0000-0002-4999-7476 (E. H. Haeusler); 0000-0003-3073-3734 (S. Lifschitz)

Structure. Section 2 introduces the foundations and design choices; Section 3 reviews related work; Section 4 describes the architecture and experimental setup; Section 5 presents results and analysis; Section 6 concludes and outlines future work.

2. Foundations

Database tuning is a critical task to enhance the performance of relational database systems. It involves selecting and applying actions—such as index creation, parameter adjustment, and query rewriting—to reduce execution time and resource consumption. Traditional tuning strategies often rely on manual decisions based on the DBA's experience or automated tools that behave as black boxes. These approaches lack transparency, reproducibility, and adaptability, especially in dynamic and heterogeneous workloads.

To address these challenges, the OnDBTuning provides a formal and extensible model of the tuning domain, including concepts such as tables, columns, queries, indexes (simple, composite, partial), materialized views, and cost-based decision rules. OnDBTuning adopts Semantic Web standards such as RDF and OWL for knowledge representation. SPARQL is employed not only to query RDF data but also to encode inference rules that capture tuning heuristics, such as the automatic creation of indexes and materialized views [2].

Initially, the inference mechanism of OnDBTuning relied on TopSPIN, a proprietary SPIN engine integrated into TopBraid Composer [6]. However, these dependencies were removed due to licensing constraints and the need for broader applicability. Proprietary functions like `spif:split`, `spif:countMatches`, `spif:trim`, and `spif:localName` were rewritten using standard SPARQL 1.1 constructs or replaced with equivalent custom logic. This shift enabled the use of open technologies such as Apache Jena [7] and the SPIN RDF vocabulary directly, without proprietary extensions.

Complementing the ontology, the OuterTuning framework is a Java-based application that offers a flexible environment for simulating, monitoring, and evaluating the impact of tuning actions. It consists of components such as a workload analyzer, rule execution engine, graphical interface, and simulation orchestrator. OuterTuning provides visual feedback on tuning suggestions, enabling users to compare different scenarios using real cost estimates interactively.

Originally, OuterTuning employed hardcoded SQL heuristics, limiting its extensibility. Integrating with OnDBTuning, the system was restructured to support semantic reasoning and knowledge-driven decisions. A RESTful Web API was implemented to automate the metadata and SQL workloads transformation into RDF, allowing OuterTuning to act as a consumer of ontology-based recommendations. This architectural improvement promotes modularity, maintainability, and compatibility with other Semantic Web applications.

OnDBTuning and OuterTuning offer a robust, explainable, and portable solution for database tuning. Their integration provides an operational benefit in terms of performance optimization and is a didactic tool for teaching database administration and semantic web reasoning.

3. Related Work

The state of the art in relational database tuning includes both commercial tools and academic approaches. Commercial solutions such as the PostgreSQL Workload Analyzer (PoWA) offer monitoring and tuning support based on system statistics and query performance [8]. While useful for performance inspection and index recommendation, these tools are often limited to specific environments, lack formal semantic reasoning, and provide limited transparency, making them inadequate for pedagogical use or research extensibility.

Academic efforts increasingly explore advanced optimization techniques, including Bayesian optimization, reinforcement learning, and, more recently, large language models (LLMs). These approaches aim to dynamically automate and adapt tuning actions, often by learning from historical workloads.

Bayesian Optimization. Tools such as OtterTune and ResTune apply Bayesian optimization to automatically adjust configuration parameters in relational databases [9][10]. OtterTune, for example, uses Gaussian process models to recommend configurations based on historical workload data, while ResTune extends this idea with constrained optimization techniques to address latency and throughput requirements.

Reinforcement Learning. SmartIX and rCOREIL represent efforts to apply reinforcement learning (RL) to the automatic tuning of indexes [11][12]. In these systems, an RL agent explores different index configurations and learns which ones yield better performance over time. These approaches have shown strong potential in dynamic and heterogeneous workloads but remain complex to interpret and challenging to integrate with human-readable reasoning mechanisms.

Semantic Approaches. While LLMs and ML-based systems aim for adaptivity, they often lack transparency. Semantic Web approaches, including OnDBTuning, fill this gap by encoding expert knowledge in a formal ontology, enabling rule-based inference through SPARQL. Although fewer in number, such systems offer interpretable and customizable alternatives to opaque ML pipelines.

In contrast to previous ontology-based efforts that focused on schema mapping or query reformulation, OnDBTuning is specifically designed to capture fine-tuning strategies. Combined with OuterTuning—a Java-based tool for simulating and visualizing tuning actions—the integrated solution enables an open, extensible, and explainable workflow that bridges theoretical reasoning and practical experimentation.

4. The Experiments

We implemented a complete experimental setup combining semantic inference with real-time workload analysis to validate the proposed integration between the OnDBTuning ontology and the OuterTuning framework[1]. This integration was achieved by extending the architecture of OuterTuning with a RESTful Web API, which is responsible for mediating communication between the framework and the inference engine. The API receives a JSON payload containing metadata such as table structures and executed SQL queries, and returns a list of tuning recommendations including rule identifiers, associated queries, expected benefit (bonus), and the corresponding SQL command.

The integration followed the Strangler Pattern [13], initially allowing both legacy and new components to run in parallel. This ensured a safe migration by comparing inference outputs and validating their consistency before deactivating the static logic previously embedded in the system. The API architecture allows for modularity and extensibility, allowing it to incorporate other ontologies or inference mechanisms in the future.

Figure 1 presents the expanded architecture of the OuterTuning framework after integration with OnDBTuning.

The architecture is composed of the following components:

Components. As shown in Fig. 1, (1) *Database* — JDBC access for workload monitoring and action execution; (2) *OuterTuning Framework* — orchestration layer; (3) *Function Libraries* — extract workload features; (4) *OnDBTuning Ontology* — domain concepts and SPARQL rules; (5) *Web API* — JSON↔RDF mediation and suggestion delivery; (6) *Interface* — DBA inspection/validation; (7) *JDBC Connections* — DB interaction; (8) *WorkloadCollector* — queries, plans and frequencies; (9) *FunctionExecutor* — applies libraries to features; (10) *ConceptInstantiator* — creates RDF instances; (11) *SemanticReasoner* — rule execution over the ontology; (12) *TuningActionExecutor* — applies actions (auto or user-guided); (13) *Execution Feedback* — measures outcomes for iterative refinement.

We use the TPC-H benchmark to evaluate the framework's effectiveness, a well-established decision support workload [14]. We generated 660 queries by creating 30 variants of each of the 22 standard TPC-H queries, covering a wide range of complexity, including multiple joins, filters, and aggregations. These queries were used to simulate realistic database usage scenarios and to test the framework's capability to extract metadata, instantiate concepts, and infer tuning actions.

[1] Source code available at https://github.com/EricRLeao1311/outer_tuning/tree/outertuning_expandido

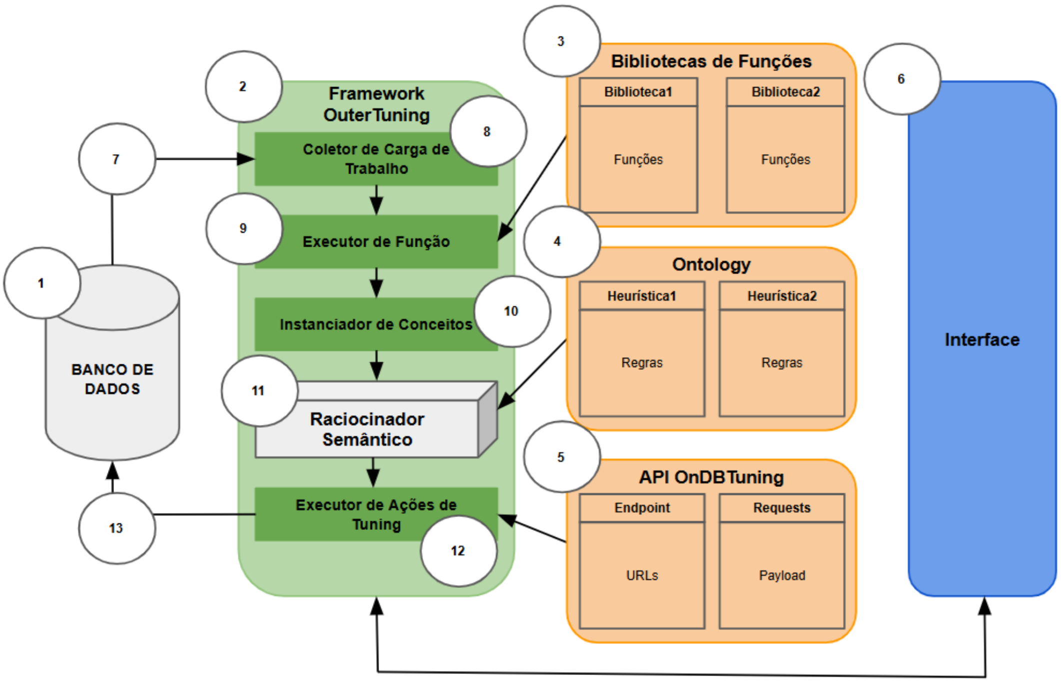


Figure 1: Expanded architecture of the OuterTuning framework

Four rule types were evaluated: *RuleHypSimpleIndex* (for suggesting single-column indexes) , *RuleHypCompositeIndex* (for generating composite indexes that combine multiple columns), *RuleSimplePartialIndex* (for creating partial indexes with specific WHERE clause filters), and *RuleHypViewAdapted* (for generating materialized views based on common aggregations). The evaluation included metrics such as the number of times each rule was triggered, the average bonus per action, and the number of queries each suggested tuning action benefited. Results showed that index-based rules—straightforward and composite indexes—provided broader impact, while materialized views and partial indexes were more specific and restrictive in their applicability.

Overall, the experimental integration proved the feasibility and effectiveness of combining declarative semantic reasoning with a workload-aware tuning framework. The modular API design and inference pipeline allow future extension and promote the adoption of explainable optimization mechanisms in both academic and professional database tuning scenarios.

5. Results

To assess the effectiveness of the integrated framework, we conducted a comparative performance evaluation using the TPC-H benchmark with a scale factor of 0.01 (approximately 10MB). The evaluation considered both the quality of tuning suggestions and their impact on system performance under different index configurations.

5.1. Evaluation Metrics

We adopted standard TPC-H performance metrics to quantify the benefits of each tuning strategy:

- **Power:** Measures performance under a single-stream workload (sequential query execution). It is computed as:

$$\text{Power} = \frac{3600 \times \text{SF}}{\left(\prod_{i=1}^{n} t_i\right)^{1/n}}$$

where t_i represents the execution time of each operation (queries and refresh functions), and SF is the scale factor.

- **Throughput:** Measures performance under multi-stream execution (parallel queries). It is calculated as:

$$\text{Throughput} = \frac{S \times N \times 3600 \times \text{SF}}{T}$$

where S is the number of streams, N the number of operations per stream, and T the total elapsed time.
- **QphH (Queries per hour):** The geometric mean of Power and Throughput:

$$\text{QphH} = \sqrt{\text{Power} \times \text{Throughput}}$$

- **Index Size:** Total disk space consumed by all indexes, measured in megabytes.

5.2. Compared Configurations

We compared the following tuning strategies:

- **No Index:** Baseline with no indexes applied.
- **DBA Expert:** Manually selected indexes simulating expert knowledge.
- **POWA:** Automatic recommendations using PostgreSQL Workload Analyzer.
- **rCOREIL:** Reinforcement learning-based tuning with dynamic index combinations.
- **OuterTuning Simple / Complete:** configurations applying either only simple indexes (Simple) or the full set of simple plus composite indexes (Complete) as our framework recommends.
- **OuterTuning Simple Selected / Complete Selected:** Subsets of indexes manually validated using the GUI to balance performance and index size.

5.3. Results and Analysis

Table 1 presents the evaluation results for each configuration.

Table 1
performance of tuning strategies with TPC-H (SF=0.01)

Configuration	Power	Throughput	QphH	Index Size (MB)
OuterTuning Complete Selected	**2013.96**	**1398.81**	**1678.44**	**4.93**
OuterTuning Simple Selected	1693.14	724.48	1107.54	2.71
rCOREIL	1995.72	547.45	1045.25	20.22
OuterTuning Simple	1432.78	443.78	797.40	7.00
OuterTuning Complete	1811.23	257.70	683.19	53.31
DBA Expert	1374.61	172.34	486.73	1.58
POWA	1219.53	159.37	440.86	2.07
No Index	1007.00	178.96	424.51	0.00

The *OuterTuning Complete Selected* configuration outperformed all others, achieving a QphH of 1678.44 — a 295% improvement over the *No Index* baseline. It also maintained a compact index footprint of only 4.93 MB, illustrating that selective application of high-impact indexes via the graphical interface leads to an optimal trade-off between performance and storage cost.

Although *rCOREIL* achieved a comparable Power score, its QphH was 38% lower than that of *OuterTuning Complete Selected*, while using over four times the index size. This underscores the effectiveness of explainable, rule-based inference combined with human-in-the-loop filtering.

The *OuterTuning Complete* configuration, which indiscriminately applied all suggested indexes, showed worse results than its selected variant. This reinforces the importance of prioritizing quality over quantity in index selection. Similarly, traditional strategies like *POWA* and *DBA Expert* delivered lower QphH scores due to their reliance on simple indexes and lack of adaptive reasoning.

In summary, the results validate the advantages of the proposed framework in achieving high performance while maintaining interpretability, modularity, and efficient resource usage.

6. Conclusion

This work presented the integration of the OnDBTuning ontology into the OuterTuning framework, resulting in a semantic, extensible, and explainable system for semi-automatic tuning of relational databases. By combining rule-based reasoning with real-time workload monitoring, the integrated framework can generate interpretable tuning recommendations — such as indexes and materialized views — and apply them in a modular, user-guided manner.

The proposed Web API and its decoupled architecture enabled seamless integration with OuterTuning, allowing dynamic inference using SPARQL and eliminating dependencies on proprietary tools. The results obtained through the TPC-H benchmark demonstrated that the *OuterTuning Complete Selected* configuration achieved superior performance in both Power and Throughput, while maintaining low storage overhead. Compared to traditional tools such as POWA and reinforcement learning approaches like rCOREIL, our approach provided competitive — and in many cases superior — results, with greater transparency and flexibility.

However, despite the promising findings, future work must address some limitations to ensure robustness and generalizability. First, it is necessary to compare the system with a broader set of tuning tools that support simple indexes, composite indexes, and materialized views. This will allow for a more complete and balanced evaluation of the framework's inference capabilities.

Second, the current experiments were conducted using a single database instance with a scale factor of 0.01 (approximately 10MB), and the rCOREIL results were taken from its published paper. A more rigorous evaluation should involve running both tools under the same conditions and on diverse databases with varying workloads and sizes. This would help ensure fair and unbiased comparisons and confirm the scalability and adaptability of the proposed solution.

As future work, we plan to address these gaps by extending the benchmarking suite, increasing data scale, and expanding the range of inference strategies integrated into the framework to strengthen its reliability and practical applicability in real-world database environments.

Acknowledgments

This study was partially funded by the CAPES Institutional Grant for PUC-Rio. Sergio Lifschitz is partially funded by CNPq Research Grant 313504/2025-3, Edward H. Haeusler is partially funded by CNPq Research Grant 309287/2023-5, and FAPERJ APQ1 Grant E-26/210.258/2019.249292. The author is supported by a CAPES scholarship and gratefully acknowledges this support.

Declaration on Generative AI

While preparing this work, the authors used GPT-4o to: Grammar, spelling check, and Text translation. After using this tool, the authors reviewed and edited the content as needed and assume full responsibility for the publication's content.

References

[1] D. Shasha, P. Bonnet, Database Tuning: Principles, Experiments and Troubleshooting Techniques, Morgan Kaufmann, San Francisco, 2002.

[2] A. C. Almeida, M. L. M. Campos, F. Baião, S. Lifschitz, R. P. de Oliveira, D. Schwabe, An ontological perspective for database tuning heuristics, in: A. H. F. Laender, B. Pernici, E. Lim, J. P. M. de Oliveira (Eds.), Conceptual Modeling - 38th International Conference, ER 2019, Salvador, Brazil, Proceedings, volume 11788 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 240–254.

[3] S. Staab, R. Studer, Handbook on Ontologies, Springer, 2010.

[4] L. de Sá Silva Perciliano, V. dos Santos, F. Baião, E. H. Haeusler, S. Lifschitz, A. C. Almeida, Inferencing Relational Database Tuning Actions with OnDBTuning Ontology, in: Procs Brazilian Symposium on Databases, SBBD, Rio de Janeiro, Brazil (Online), 2021, pp. 157–168.

[5] R. P. de Oliveira, F. Baião, A. C. Almeida, D. Schwabe, S. Lifschitz, Outer-tuning: an integration of rules, ontology and RDBMS, in: F. G. Rocha, I. Vasconcelos, R. P. dos Santos, D. Viana, S. de Avila e Silva (Eds.), Proceedings of the XV Brazilian Symposium on Information Systems, SBSI, Aracaju, Brazil, 2019, pp. 60:1–60:8.

[6] SPIN RDF, Spin – sparql inferencing notation, 2024. https://spinrdf.org/.

[7] Apache Jena, Apache jena documentation, 2024. https://jena.apache.org/.

[8] Dalibo, PoWA: PostgreSQL Workload Analyzer — documentation, https://powa.readthedocs.io/, 2025. Version 5.0.0, accessed 2025-07-13.

[9] D. V. Aken, A. Pavlo, G. J. Gordon, B. Zhang, Automatic database management system tuning through large-scale machine learning, in: S. Salihoglu, W. Zhou, R. Chirkova, J. Yang, D. Suciu (Eds.), Proceedings of the 2017 ACM International Conference on Management of Data SIGMOD, Chicago, IL, USA, 2017, pp. 1009–1024.

[10] X. Zhang, H. Wu, Z. Chang, S. Jin, J. Tan, F. Li, T. Zhang, B. Cui, Restune: Resource oriented tuning boosted by meta-learning for cloud databases, in: G. Li, Z. Li, S. Idreos, D. Srivastava (Eds.), SIGMOD '21: International Conference on Management of Data, Virtual Event, China, 2021, pp. 2102–2114.

[11] G. P. Licks, J. M. C. Couto, P. de Fátima Miehe, R. D. Paris, D. D. A. Ruiz, F. Meneguzzi, Smartix: A database indexing agent based on reinforcement learning, Applied Intelligence 50 (2020) 2575–2588.

[12] D. Basu, Q. Lin, W. Chen, H. T. Vo, Z. Yuan, P. Senellart, S. Bressan, Regularized cost-model oblivious database tuning with reinforcement learning, Transactions on Large Scale Data Knowledge Centered Systems 28 (2016) 96–132.

[13] S. Newman, Monolith to Microservices: Evolutionary Patterns to Transform Your Monolith, O'Reilly Media, 2020.

[14] Transaction Processing Performance Council, TPC-H benchmark specification, revision 2.17.3, 2023. https://www.tpc.org/tpch/.

Domain ontology for mapping competency development in higher education engineering programs

Eduardo Miguel Perotti Oliveira[1], Eduardo Ribeiro Felipe[2], Fernanda Farinelli[3], Giovani Bernardes Vitor[4] and Rodrigo Aparecido da Silva Braga[5]

[1]*Federal University of Itajubá, Institute of Science and Technology, Itabira, Minas Gerais, Brasil*
[2]*Federal University of Itajubá, Institute of Science and Technology, Itabira, Minas Gerais, Brasil*
[3]*University of Brasília, Faculty of Information Science, Brasília, Distrito Federal, Brasil*
[4]*Federal University of Itajubá, Institute of Science and Technology, Itabira, Minas Gerais, Brasil*
[5]*Federal University of Lavras, Institute of Science, Technology and Innovation, São Sebastião do Paraíso, Minas Gerais, Brasil*

Abstract

This manuscript outlines an ongoing master's research project focused on the development of a domain ontology to support the mapping and monitoring of competencies acquired by students throughout their academic programs. The methodology combines the Realism-Based Ontology Engineering Methodology (ReBORM), the Basic Formal Ontology (BFO), and the competencies outlined in the Conceive–Design–Implement–Operate (CDIO) framework. In contrast to existing approaches, this integration enables semantic traceability between courses, content, and competencies, supporting curriculum analysis and alignment with labor market expectations. The ontology supports terminological standardization, ensures interoperability across curricular structures, and provides a foundation for the automatic assessment of competencies and identify gaps in program design. Although initially applied to a computer engineering program, the ontology is designed to be extensible to other educational programs. This paper details the research context, methodology, and preliminary modeling results, with the empirical validation using actual curricular data planned for the subsequent research phase.

Keywords

ontology, competencies, computer engineering, CDIO, high education curriculum

1. Introduction

In a society accelerated by the rise of artificial intelligence, the systematization of knowledge becomes imperative. Within competency-based academic education [1], this systematization is instrumental for evaluating the curricular structures of educational institutions and aligning graduate profiles with labor market expectations [2].

Despite numerous definitions, a competency can be understood as the ability to mobilize knowledge, personal skills, and socio-methodological competencies to solve problems in educational or professional contexts [3, 4]. In engineering, for example, high-level competencies combine technical-scientific mastery with the capacity for innovation and adaptation to complex scenarios [5]. Several efforts to map the competencies acquired by students throughout the curricular structure can be found in the literature [6].

In line with this perspective, the notion of mobilizing knowledge related to competencies is also a cornerstone of contemporary discussions on information literacy, a concept widely discussed in the literature, and often defined as the ability to locate, evaluate, and use information effectively [7]. A critical approach to this concept, critical information literacy, expands this definition to include the capacity to mobilize knowledge and use it to act upon complex problems and question the power structures embedded in the production and dissemination of information.

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

emperotti@unifei.edu.br (E. M. P. Oliveira); eduardo.felipe@unifei.edu.br (E. R. Felipe); fernanda.farinelli@unb.br (F. Farinelli); giovanibernardes@unifei.edu.br (G. B. Vitor); rodrigobraga@ufla.br (R. A. d. S. Braga)
0009-0008-7755-1501 (E. M. P. Oliveira); 0000-0003-1690-2044 (E. R. Felipe); 0000-0003-2338-8872 (F. Farinelli); 0000-0002-5807-5465 (G. B. Vitor); 0000-0002-5490-9944 (R. A. d. S. Braga)

The systematization of competencies in the field of engineering has spurred the proposition of various theoretical and methodological models [8]. It is important to note, however, that competencies are not restricted to technical or engineering domains. They also encompass interpersonal abilities such as conflict mediation, synthesis of perspectives, and consensus building. Prominent among these initiatives is the CDIO (Conceive, Design, Implement, Operate) model, which is internationally established as a reference for the integration of technical and transversal skills throughout the lifecycle of engineering projects, from conception to operation [9].

However, despite these efforts, current approaches still reveal important limitations. Pedagogical frameworks such as CDIO offer structured competency guidelines but lack formal semantic representations to ensure interoperability and traceability, while ontological initiatives in education often remain disconnected from pedagogical foundations, which restricts their applicability in practice. As a result, there is still no systematic, semantically rigorous, and pedagogically grounded model capable of effectively supporting the monitoring of student learning outcomes. This gap is particularly critical in undergraduate education, where, in addition to technical expertise, transversal competencies such as teamwork, leadership, creativity, and information literacy must be systematically developed and assessed to align graduate profiles with societal and labor market demands.

In this context, ontologies emerge as a tool capable of addressing the need for terminological standardization and semantic relationships in the management of graduate attributes [3]. The development of specialized ontologies can unify competency frameworks and mitigate conceptual ambiguities, thereby promoting greater interoperability in the integration of educational systems and models.

The objective of this work is to develop an ontological representation that enables the tracking of competencies acquired by students throughout the educational path defined in the curricular structures of higher education programs. Although initially applied to a computer engineering program, the proposed model seeks sufficient generality and flexibility to be adapted to different knowledge areas and academic contexts. This allows for the analysis and monitoring of educational development in programs of diverse natures and complexity levels.

Thus, the central question the ontology seeks to address is how the competencies of an undergraduate program are distributed across its curricular structure. To achieve this objective and answer this question, the Realism-Based Ontology Engineering Methodology (ReBORM) and the Basic Formal Ontology (BFO) will be employed, in conjunction with the competencies delineated in the Conceive-Design-Implement-Operate (CDIO) framework. To support the development and validation of the proposed ontology, this research also adopts the Design Science Research (DSR) [10] methodology, which provides a structured framework for the creation and evaluation of innovative artifacts.

This article is structured as follows: Section 2 presents the theoretical background, covering concepts of ontologies and competency mapping based on the CDIO model. Section 3 reviews related works that contextualize this research. Section 4 describes the adopted methodology. Section 5 presents the ontology modeling and discusses the results obtained. Finally, Section 6 provides the final considerations.

2. Theoretical background

2.1. Ontologies

An ontology, in the context of computer science and the field of knowledge representation, is defined as an explicit, formal specification of a shared conceptualization[11]. As a representational artifact, an ontology's primary characteristics are: (1) formalization, which entails the use of logic and standardized languages to ensure precision; (2) conceptualization, which organizes domains into concepts, relations, and axioms; (3) sharedness, as it reflects a consensus on a domain among agents or communities; and (4) reusability, allowing its application across different contexts [12]. These characteristics underpin its utility in knowledge modeling, systems interoperability, and semantic inference.

Ontology reuse is an essential practice to ensure consistency, interoperability, and efficiency in knowledge representation, thereby avoiding redundant effort in the construction of new models [13]. In this work, the Basic Formal Ontology (BFO) is adopted as a reference ontology due to its fundamental

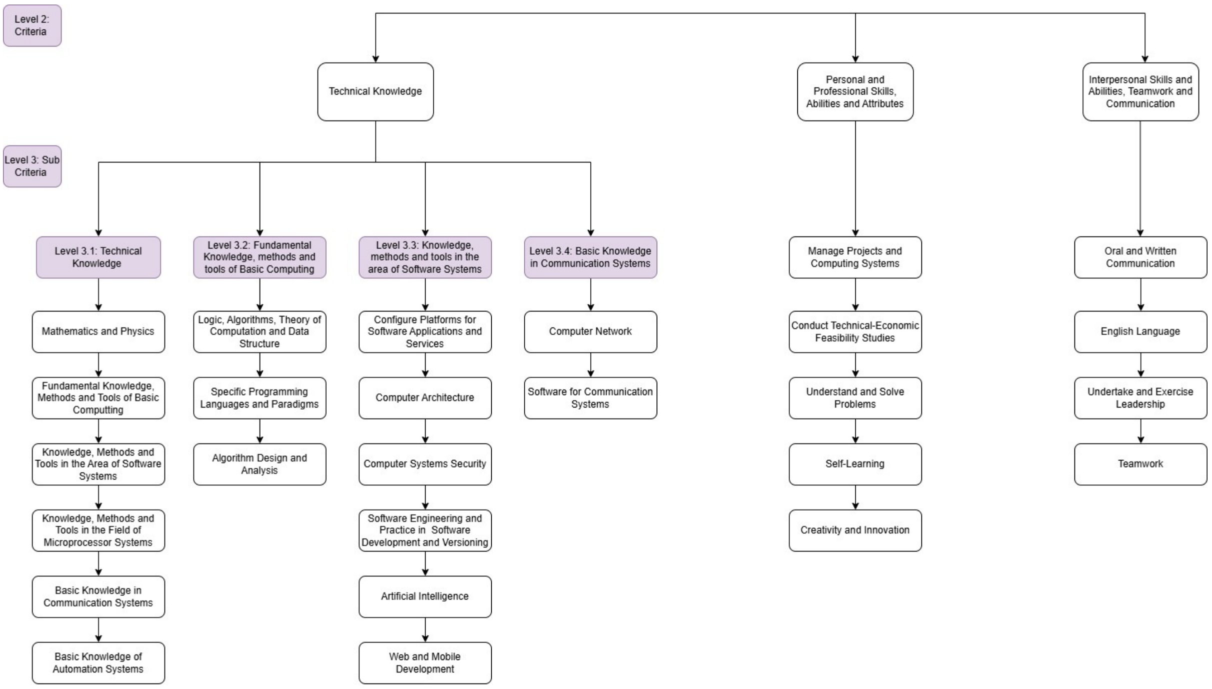

Level 2: Criteria
Technical Knowledge
Personal and Professional Skills, Abilities and Attributes
Interpersonal Skills and Abilities, Teamwork and Communication
Level 3: Sub Criteria
Level 3.1: Technical Knowledge
Level 3.2: Fundamental Knowledge, methods and tools of Basic Computing
Level 3.3: Knowledge, methods and tools in the area of Software Systems
Level 3.4: Basic Knowledge in Communication Systems
Mathematics and Physics
Fundamental Knowledge, Methods and Tools of Basic Computing
Knowledge, Methods and Tools in the Area of Software Systems
Knowledge, Methods and Tools in the Field of Microprocessor Systems
Basic Knowledge in Communication Systems
Basic Knowledge of Automation Systems
Logic, Algorithms, Theory of Computation and Data Structure
Specific Programming Languages and Paradigms
Algorithm Design and Analysis
Configure Platforms for Software Applications and Services
Computer Architecture
Computer Systems Security
Software Engineering and Practice in Software Development and Versioning
Artificial Intelligence
Web and Mobile Development
Computer Network
Software for Communication Systems
Manage Projects and Computing Systems
Conduct Technical-Economic Feasibility Studies
Understand and Solve Problems
Self-Learning
Creativity and Innovation
Oral and Written Communication
English Language
Undertake and Exercise Leadership
Teamwork

Figure 1: Expected competencies for graduates of the computer engineering program at the Federal University of Itajubá – Itabira Campus. Image by P. A. C. Santos, G. B. Vitor, R. A. d. S. Braga, A. C. O. Santos [2]

characteristics: (1) philosophical rigor [14]; (2) broad recognition [15]; and (3) interoperability [16]. BFO was selected for its capacity to provide a top-level framework that supports the coherent organization of entities across diverse domains, facilitating both modeling and reuse.

To operationalize the structured integration and reuse of ontologies, Farinelli et al. (2017) [17, 18] propose the ReBORM (Realism-Based Ontology engineering Methodology), which combines principles of ontological realism [19] with practices from the NeOn methodology [20]. According to Farinelli et al. (2017), ontological realism aligns with the philosophical rigor of BFO by emphasizing the precise demarcation of the domain and the correspondence between ontological entities and reality. The NeOn methodology, in turn, provides an iterative-incremental cycle organized into phases—such as conceptualization, inception, design, implementation, and delivery—ensuring that the developed ontology remains aligned with scientific principles and engineering praxis [18].

2.2. CDIO and competency mapping

The CDIO is an educational framework widely adopted in engineering education around the world, focusing on the development and planning of competency- and outcome-based curricula [21]. The CDIO Syllabus [21] outlines a set of knowledge, skills, and attitudes considered desirable for students, and it is flexible enough to be adapted by any engineering education institution.

At the university under study, version 2.0 of the CDIO Syllabus is adopted. Figure 1 presents the competencies defined in the course's pedagogical project, in accordance with the principles established by CDIO 2.0.

3. Related work

To address the gap between the qualifications offered by vocational education and the competencies required by the labor market, [3] details the development of job-Know. This is an ontology designed to forge an explicit link between the knowledge, skills, and abilities (KSAs) developed in vocational education and training (VET) and the competency prerequisites of job positions. The conceptual basis for this ontology is a three-dimensional semantic framework (TCK 3-D), which models the interrelationships among task, competence, and knowledge. Thus, job-know provides a formal and computationally processable representation of the work and education domains, enabling systematic analysis and alignment between what is taught and what is required.

To overcome the challenges of hierarchical alignment between macro-level educational objectives and micro-competencies in outcome-based education (OBE) curricula, [5] proposes OBC-ONTO, an ontology that models vertical (program → course) and horizontal (interdisciplinary) coherence in higher education. This framework formalizes the relationship among program educational objectives (PEOs), program learning outcomes (PLOs), and course learning outcomes (CLOs), utilizing a learning experience matrix that links each PLO to specific pedagogical activities, content, technologies, and assessments. Unlike previous models, which omitted the representation of program-level outcomes, this ontology offers critical support for curriculum reviewers and accreditation processes, ensuring that declared competencies are traceable, measurable, and aligned with industry demands.

Meanwhile, [22] proposes a new paradigm for engineering education, centered on the integration of projects and competencies. Its objective is to establish an ontology-based knowledge representation model, formally mapping professional domains to curricular structures. To this end, it defines hierarchical competency models (bachelor's, master's, doctorate), where declarative knowledge is decomposed into detailed ontologies — encompassing core concepts, identifiers, and concretizers — which serve as a basis for curriculum construction.

Considering the related works presented, this manuscript addresses a gap in the literature: the lack of interaction between pedagogical frameworks and ontologies for mapping and assessing competencies in Engineering programs. Specifically, no ontological representation integrated with the CDIO framework was found. Thus, this manuscript proposes to fill this gap by means of an ontology implemented in Protégé, which: (i) structures technical, personal, and interpersonal competencies aligned with the CDIO framework; (ii) employs inference rules for the automatic assessment of acquired skills; and (iii) is validated using real data from the program at the Federal University of Itajubá (Unifei), Itabira campus.

4. Methodology

As mentioned in Section 2.1, this work adopts the ReBORM methodology for ontology design, structured into five phases: 1) concept, 2) inception, 3) design, 4) implementation, and 5) delivery. Therefore, the competency mapping for the computer engineering program, based on the CDIO framework[2] and presented in Fig. 1, will be represented in the BFO ontology following the phases defined by the ReBORM methodology.

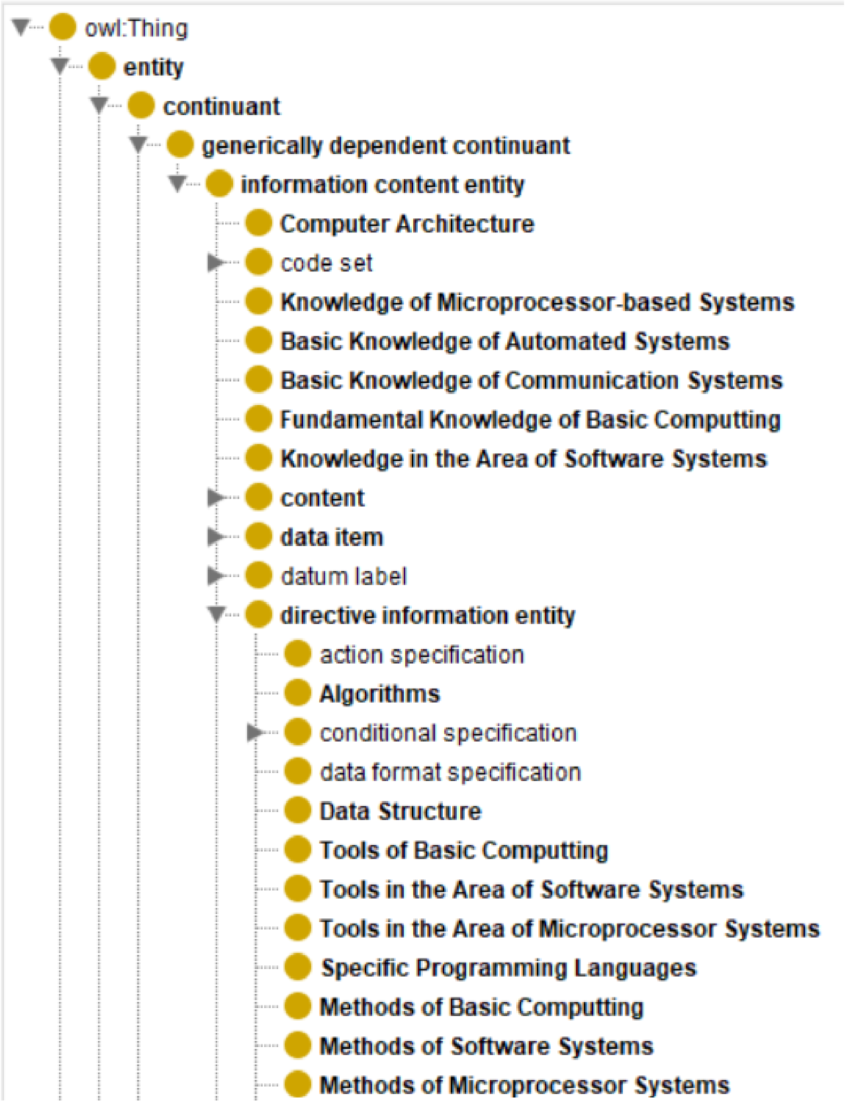
This research adopts the Design Science Research (DSR) methodology [10], a paradigm suited for the development and validation of innovative artifacts designed to solve identified problems. The DSR cycle guides this work through three core iterative phases: (i) problem identification, which motivates the need for the artifact; (ii) artifact development, which encompasses the design and construction of the solution; and (iii) evaluation, which assesses the artifact's utility and efficacy in addressing the problem.

As a result of phases 1 and 2, the Ontology Requirements Specification Document (OSRD) is available at [23].

The competencies identified in Fig. 1 were decomposed into terms[24]. Subsequently, by applying steps (3) and (4) of the ReBORN methodology, a hierarchy was constructed based on the Basic Formal Ontology (BFO). The final ontology file, corresponding to phase 5, is available at [25]. Section 5 presents the results and a discussion of the ontology.

5. Ontology modeling and discussion

The ontology proposed in this study, hereafter referred to as CompOnt, is partially illustrated in Fig. 2. A consolidated summary of all competencies as ontological entities, along with their respective classifications, can be found in [24].



- owl:Thing
  - entity
    - continuant
      - generically dependent continuant
        - information content entity
          - Computer Architecture
          - code set
          - Knowledge of Microprocessor-based Systems
          - Basic Knowledge of Automated Systems
          - Basic Knowledge of Communication Systems
          - Fundamental Knowledge of Basic Computting
          - Knowledge in the Area of Software Systems
          - content
          - data item
          - datum label
          - directive information entity
            - action specification
            - Algorithms
            - conditional specification
            - data format specification
            - Data Structure
            - Tools of Basic Computting
            - Tools in the Area of Software Systems
            - Tools in the Area of Microprocessor Systems
            - Specific Programming Languages
            - Methods of Basic Computting
            - Methods of Software Systems
            - Methods of Microprocessor Systems

Figure 2: Section of CompOnt as represented in Protégé. Image by Authors [25]

In contrast, informational artifacts that serve as a plan or set of instructions to guide a process were categorized as directive information entity (**continuant - generically dependent continuant - directive information entity**). Such entities have a prescriptive nature. This group includes **algorithms**, **data structure**, **specific programming languages**, **programming paradigms**, **softwares for communication systems**, as well as the **methods** and **tools** associated with **basic computing** and the **area of software systems** and **microprocessor**.

Activities and actions that unfold over time and are executed according to a plan, method, or design were mapped to the planned process class, located in the **ocurrent - process** branch. This category encompasses complex processes such as **understand and solve problems**, **conduct technical economic feasibility**, **algorithms analysis and design**, and **software development**—including its **web** and **mobile** specializations. Also classified in this category were operational processes such as **configure platforms for software applications**, the execution of **software services**, and **software versioning**.

Transversal competencies and abilities intrinsic to an agent (e.g., a student or professional) were classified as quality (**continuant - specifically dependent continuant - quality**). Such entities are inherent in and dependent on a specific bearer for their existence. Examples include **self-learning**, **creativity** and **innovation**, **oral and written communication**, **leadership**, and **teamwork**.

Finally, entities that represent material objects and physical systems, which exist independently in spacetime, were categorized as object aggregate (**continuant - independent continuant - material entity - object aggregate**). The concepts of **computer network** and **computer systems** were allocated to this class, as they represent tangible sets of hardware components.

6. Final considerations

This work presented an ongoing research effort toward the development of a domain ontology for mapping competencies in higher education programs. The proposed model integrates ReBORM, BFO, and CDIO, and has been preliminarily applied to the computer engineering program at UNIFEI. The results demonstrate the feasibility of formally representing both technical and transversal competencies, enabling interoperability and potential reuse across different educational contexts.

The proposed ontology differs from existing solutions by providing a unified semantic framework that bridges pedagogical foundations (CDIO) with ontological rigor (BFO/ReBORM). This integration allows for precise competency tracing across curricular components, supporting gap analysis and alignment with labor market requirements. By enabling automated assessment and curriculum evaluation, the ontology offers a practical approach to developing professional profiles that better meet societal and industry needs.

However, the research is still in progress. The empirical validation of the ontology, through its application to real curricular data and stakeholder evaluation, remains a future stage. This will be crucial for confirming its effectiveness as a tool for curriculum mapping, competency assessment, and support for pedagogical decision-making.

By systematically integrating pedagogical frameworks and reference ontologies, the proposed approach contributes to overcoming current limitations in competency-based education. The primary technological contribution is the CompOnt ontology, which provides a reusable and extensible semantic framework for diverse academic areas. This extensibility ensures the ontology can evolve to incorporate new courses, institutions, and curricular structures, thereby advancing the state of the practice in competency mapping.

7. Declaration on Generative AI

During the preparation of this work, the author(s) used **DeepSeek** in order to: **identify and correct grammatical errors**, **typos**, **and other writing mistakes**, **rephrase sentences or paragraphs to improve clarity**, **conciseness**, **or style**. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] G. Le Boterf, De la compétence à la navigation professionnelle (1999).

[2] P. A. C. Santos, G. B. Vitor, R. A. d. S. Braga, A. C. O. Santos, Engineering student profile related to the market of operation: Competence perspective, IEEE Transactions on Education 66 (2022) 45–54.

[3] A. Atif, P. Busch, D. Richards, Towards an ontology-based approach to knowledge management of graduate attributes in higher education, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2012). doi:10.1007/978-3-642-32541-0_20.

[4] A. Kalou, G. Solomou, C. Pierrakeas, A. Kameas, An ontology model for building, classifying and using learning outcomes, Proceedings of the 12th IEEE International Conference on Advanced Learning Technologies, ICALT 2012 (2012) 61 – 65. doi:10.1109/ICALT.2012.45.

[5] S. Aminah, A. Alfa Krisnadhi, A. Nizar Hidayanto, Ontological framework for the analysis of outcome-based curriculum in higher education, IEEE Access (2025). doi:10.1109/ACCESS.2025.3542881.

[6] M. Abelha, S. Fernandes, D. Mesquita, F. Seabra, A. T. Ferreira-Oliveira, Graduate employability and competence development in higher education—a systematic literature review using prisma, Sustainability 12 (2020) 5900.

[7] E. Tewell, A decade of critical information literacy: A review of the literature, Communications in Information Literacy 9 (2015) 24–43. URL: http://archives.pdx.edu/ds/psu/22378. doi:10.15760/comminfolit.2015.9.1.174.

[8] A.-K. Winkens, F. Engelhardt, C. Leicht-Scholten, Resilience-related competencies in engineering education–mapping abet, eur-ace and cdio criteria (2023).

[9] F. Cheng, Reform and practice for cdio engineering education model in curriculum teaching of mechanical design of higher vocational education of china, Advanced Materials Research (2011). doi:10.4028/www.scientific.net/AMR.271-273.1228.

[10] A. Dresch, D. P. Lacerda, J. A. V. A. Júnior, Design Science Research: A Method for Science and Technology Advancement, Springer International Publishing, Cham, 2015. doi:10.1007/978-3-319-07374-3.

[11] T. Gruber, A translational approach to portable ontologies, Knowledge Acquisition 5 (1993) 199–220. doi:10.1006/knac.1993.1008.

[12] N. Guarino, Formal ontologies and information systems (1998).

[13] R. Arp, B. Smith, A. D. Spear, Building Ontologies with Basic Formal Ontology, MIT Press, 2015.

[14] B. Smith, On classifying material entities in basic formal ontology, Interdisciplinary Ontology 5 (2012) 1–13.

[15] ISO/IEC, Information technology — top-level ontologies (tlo) — part 1: Requirements, 2021.

[16] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al., The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nature Biotechnology 25 (2007) 1251–1255.

[17] M. Barcellos Almeida, F. Farinelli, Ontologies for the representation of electronic medical records: The obstetric and neonatal ontology, Journal of the Association for Information Science and Technology 68 (2017) 2529–2542.

[18] F. Farinelli, Improving Semantic Interoperability in the Obstetric and Neonatal Domain Through an Approach Based on Ontological Realism, Ph.D. thesis, UFMG, Belo Horizonte, 2017.

[19] K. Munn, B. Smith, Applied Ontology: An Introduction, Ontos Verlag, Frankfurt, 2008.

[20] M. C. Suárez-Figueroa, NeOn Methodology for Building Ontologies, Universidad Politécnica de Madrid, Madrid, 2010.

[21] CDIO Initiative, Cdio syllabus topical form, 2024. URL: https://www.cdio.org/benefits-cdio/cdio-syllabus/cdio-syllabus-topical-form, accessed: 2025-06-05.

[22] B. Kubekov, L. Bobrov, E. Savelyeva, V. Naumenko, A. Utegenova, Project-competent paradigm of knowledge representation of the three-level engineering education system, in: CEUR Workshop Proceedings, volume 2344, 2018. URL: http://ceur-ws.org/Vol-2344/paper12.pdf.

[23] Eduardo M. Perotti, Ontology Requirements Specification Document (OSRD) - CompOnt, https://docs.google.com/spreadsheets/d/1qpdY9RhRdXYOovpVbytzfwUC532bsaPrjhntTfOPIBk/edit?usp=sharing, 2025. Accessed: 2025-06-11.

[24] Eduardo M. Perotti, Ontology CompOnt - Term Classification, https://docs.google.com/spreadsheets/d/1qpdY9RhRdXYOovpVbytzfwUC532bsaPrjhntTfOPIBk/edit?usp=sharing, 2025. Accessed: 2025-06-11.

[25] Eduardo M. Perotti, Ontology - CompOnt, https://github.com/emperotti/CompOnt, 2025. Accessed: 2025-06-11.

An ontology-based approach to streamline the reconstruction of genome-scale metabolic models

Nahim Alves de Souza[1], Renata Wassermann[1]

[1]*Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010, Cidade Universitária, São Paulo, SP, Brasil*

Abstract

The reconstruction of genome-scale metabolic models (GEMs) is a complex and laborious process that depends significantly on expert manual curation. It involves integrating data from diverse sources, such as biochemical databases and scientific literature, which often contain inconsistencies due to the lack of standardized representations for metabolites and reactions. Since current solutions cannot fully resolve these discrepancies, domain experts have to manually identify and correct them, which is a time-consuming task. This paper proposes an ontology-based approach to streamline the reconstruction of GEMs. This work proposes developing a GEM ontology to formally represent both GEM structures and the expert knowledge used during reconstruction. In the future, this ontology can be integrated into the reconstruction workflow through an application that takes draft models as input and produces enhanced models by incorporating additional information from biochemical datasets. This approach is expected to reduce manual curation effort and consequently simplify the overall reconstruction process.

Keywords

Genome-scale metabolic models, Ontology, Semantic interoperability, Data integration

1. Introduction

Metabolism is defined as a series of chemical reactions that occur continuously within living organisms to sustain life, particularly those associated with energy production and growth [1]. Nevertheless, mapping the entire metabolism of an organism is a highly complex task due to the vast number of compounds and reactions involved. The process of building a computational representation of an organism's metabolism is known as **reconstruction**, while the resulting model is often referred to as a **genome-scale metabolic model (GEM)** – since it is based on organism's genome.

Over the years, several computational resources have been developed to support the reconstruction of metabolic models, including databases, tools, and ontologies. However, these resources present significant limitations, especially in unifying data from multiple sources [2, 3, 4]. Moreover, none of the available ontologies (e.g., ChEBI [5], GO [6], SBO [7]) comprehensively represent GEMs or are actively applied in GEM reconstruction, as they contain little information about the complex relationships between reactions, metabolites, and genes – which are essential for integrating data during GEM reconstruction. Consequently, experts must rely on their domain knowledge to manually resolve data inconsistencies, suggesting that the necessary semantic information exists but is not explicitly represented in current databases and ontologies.

Based on these premises, this work proposes the development of a new ontology for representing GEMs with two main objectives: (1) to enable the integration and reconciliation of models across different datasets, and (2) to facilitate the quality assessment of GEMs through the use of logical inferences to identify (and potentially repair) inconsistencies in models. The following sections present the GEM reconstruction process and its main challenges, along with the proposed approach to address them.

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

nahim@usp.br (N. A. d. Souza); renata@ime.usp.br (R. Wassermann)
0009-0007-8051-558X (N. A. d. Souza); 0000-0001-8065-1433 (R. Wassermann)

2. GEM reconstruction

GEM reconstruction is a complex process that involves integrating data from diverse sources, conducting thorough literature reviews, performing manual curation, running mathematical simulations, and validating results through biological experiments. Each of these activities need to be carefully conducted in order to ensure the accuracy and quality of the resulting model. Thiele and Palsson [8] proposed a detailed five-stage protocol for the construction high-quality GEMs, summarized in Figure 1. The reconstruction process begins with the creation of a **draft model**, derived from the organism's genome and biochemical datasets. This initial model is then refined by experts, supported by computational tools that assess model quality and simulate the organism's metabolic behavior. Finally, the curated model, along with documentation of the reconstruction process, is compiled and published.

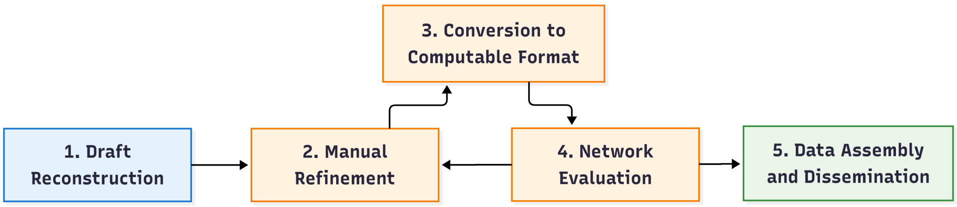

3. Conversion to Computable Format
1. Draft Reconstruction
2. Manual Refinement
4. Network Evaluation
5. Data Assembly and Dissemination

Figure 1: GEM Reconstruction stages, based Thiele and Palsson's protocol [8].

2.1. Key challenges

Integrating information on metabolites and reactions is fundamental to GEM reconstruction. Although various efforts have been made over the years to establish representation standards for biochemical data, a universal consensus has not yet been reached [4]. For instance, while the chemical structures of metabolites can be represented as SMILES strings [9], this format does not establish a unique representation for each molecule[1] [2] – water, for example, can be represented as **[OH2]**, **[H]O[H]**, or simply **O**[2]. The InChI system [10] addresses some of these ambiguities by providing a more detailed representation of molecular structures. However, certain molecules representable in SMILES may lack proper encoding in InChI, and the optional nature of some fields in the InChI format can result in incomplete representations in certain contexts [2]. Beyond variations in metabolite naming, the representation of reactions adds another layer of complexity to the reconciliation of biochemical data. The multiple ways of expressing reactions – ranging from chemical equations [3] to more complex representations such as reaction graphs, hypergraphs, or stoichiometric matrices [11, 12, 13, 14] – hinder data comparison across different datasets.

In GEMs, each reaction is associated with genes and enzymes through GPRs (gene-protein-reaction) rules, along with other attributes essential for simulating organism behavior. SBML remains the dominant format for representing GEMs, as its design facilitates model exchange, reuse, and supports simulation-related properties [15]. While its flexible design allows core capabilities to be extended, it lacks a formal specification for mandatory fields, which results in inconsistently populated fields, containing incomplete or inappropriate information [16]. Furthermore, even when the same data sources are used, GEM reconstructions can differ due to variations in methods, algorithms, and expert decisions throughout the process [17, 18].

Some approaches, such as MNXref [2] and MetRxn [3], were designed to provide a unified, reliable data source by implementing iterative reconciliation processes that utilize identifiers, names and cross-references to resolve ambiguities. Nevertheless, these methods may overlook crucial factors such as reaction directionality, compartmentalization, mass- and charge-balance, potentially leading to

[1] While Isomeric SMILES addresses some ambiguities, multiple representations remain possible [2].

[2] Although this last representation may seem unusual, it is adopted by ModelSEED (https://modelseed.org/biochem/compounds/cpd00001) and PubChem (https://pubchem.ncbi.nlm.nih.gov/compound/Water#section=SMILES)

Table 1
Competency questions were designed to reflect typical activities in GEM reconstruction. This table presents the most relevant ones.

Competency Questions
CQ01: Are metabolites *M1* and *M2* the same?
CQ02: Are reactions *R1* and *R2* the same?
CQ03: Does metabolite *M1* have the correct chemical formula?
CQ04: Is reaction *R1* mass- and charge-balanced?
CQ05: Is reaction *R* associated with a single gene or multiple genes?
CQ06: Which reactions are catalyzed by enzyme *E*?
CQ07: Is reaction *R* blocked?
CQ08: Which reactions in the network produce metabolite *M*?
CQ09: Which metabolites are dead ends in the network?
CQ10: Which genes can be knocked out to prevent the production of metabolite *M*?
CQ11: Which metabolites are produced but not consumed by any reaction in the network?
CQ12: Which metabolites are required but not produced in the network?
CQ13: Does the model generate biomass?

inaccurate results. Additionally, internal and external inconsistencies in names and identifiers, often found in biochemical databases [19], make automated data unification particularly challenging.

3. An ontology for GEMs

Ontologies have proven to be powerful tools for addressing challenges related to data integration and system interoperability [20, 21]. In the fields of chemistry and biology, numerous ontologies have been developed to establish unified vocabularies and facilitate integration across heterogeneous systems. The Basic Formal Ontology (BFO) [22], for example, is a widely adopted upper ontology that provides foundational concepts for constructing domain-specific ontologies. Notable examples include ChEBI, which models chemical entities and their relationships [5], and the Gene Ontology (GO), which represents genes along with their functions and associated products [6].

The ontology most directly associated with GEM reconstruction is the Systems Biology Ontology (SBO) [7], which defines terms related to Systems Biology, including physical entities (e.g., metabolites, genes, biomass) and processes (e.g., biochemical reactions). SBO terms can be used to annotate tags in SBML files, facilitating the integration of models originated from different sources [23]. However, despite these features, SBO does not adequately capture the complex relationships between entities, limiting its effectiveness in representing the full structure and semantics of GEMs. Furthermore, to the best of our knowledge, no existing ontology addresses this level of detail, highlighting an opportunity to overcome these challenges through the development of an ontology for GEMs, capable of enabling concept disambiguation and data integration.

3.1. Ontology development

Initially, we conducted a literature review to understand the GEM reconstruction process, existing solutions, and available databases and tools. Subsequently, in collaboration with systems biology experts, the project scope and objectives were established, and competency questions (Table 1) were formulated based on the activities typically performed during GEM reconstruction (e.g., comparing metabolites and reactions, evaluating the metabolic network). Based on these questions we identified the core concepts and relationships of the ontology and designed an initial domain model (Figure 2).

Although this model still requires refinement, it already provides a strong foundation for understanding the domain. The central component of the model is the **MetabolicModel** class, which serves as an information aggregator, encompassing properties of the organism and a list of metabolites, reactions, genes, and compartments. A **BiologicalEntity**, corresponds to entities that physically exist in the real world, which contains a unique ID, name, and a list of cross references including external data and metadata (**CrossReference** class). A **Metabolite** correspond to any molecule participating in metabolic reactions, either as a reactant, product, cofactor, or intermediate. In GEMs, each metabolite is located in a specific **Compartment** (a region within the cell, e.g., glucose in the cytosol vs. extracellular glucose are distinct). A **Reaction** is a biochemical transformation that converts a set of reactants into products. Reactions can be enzymatic (catalyzed by a single enzyme or by an enzymatic complex) or non-enzymatic (occurring spontaneously). An **Enzyme** is a special type of protein involved in the catalysis of biochemical reactions. A single enzyme may catalyze multiple reactions, and a single reaction may be catalyzed by different enzymes (the same applies to enzymatic complexes). Additionally, each enzyme can be encoded by one or more genes. A **Gene** is a DNA sequence that encodes a protein (in this context, an enzyme). Genes and reactions are linked through GPR associations, which are boolean formulas represented in the diagram through the relationships among the **Gene**, **Enzyme**, **EnzymeAssociation**, and **Reaction** classes.

In order to fully support concept disambiguation, and promote explainability and systems interoperability, the current model still requires further enhancements. For instance, several terms mentioned in the competency questions lack clear and explicit definitions, such as "same" in CQ02, "blocked reaction" in CQ07, and "required" in CQ12. Therefore, the next step, which we are currently working on, involves using the initial model as a foundation to create a more explicit model with the necessary semantic information. Recent studies have demonstrated that the application of **ontological unpacking** techniques, combined with the modeling language OntoUML, yields promising results by enriching concept definitions and making implicit knowledge explicit [24, 25, 20]. Applying these techniques to the current model can lead to a more comprehensive and semantically rich representation of GEMs.

3.2. Evaluation

The evaluation of the proposed approach must be twofold: (1) assessing the consistency and accuracy of the GEM ontology in representing domain knowledge and (2) verifying the quality of the GEMs generated using this ontology-based approach. For the first part, the representation can be assessed by encoding the ontology in OWL and using automated reasoners to detect logical inconsistencies in the knowledge base – for example, defining a reaction as occurring in one compartment while its metabolites are located in another, or incorrectly declaring two entities equivalent when they have different property values. In addition, competency questions can be translated into SPARQL queries and used to evaluate whether the ontology can answer the questions proposed by domain experts, thereby assessing the correctness and completeness of the representation [26].

The second part of the evaluation can be carried out by comparing the model generated by an application based on the GEM ontology with manually curated models from well-studied organisms, such as *Escherichia coli* (an approach commonly adopted in the literature [27, 28, 29, 21]). In this scenario, the evaluation pipeline would consist of: (1) generating a draft model from the genome of the selected organism (e.g., *E. Coli*[3]); (2) using the application to load and enhance the model; and (3) comparing the resulting model with a manually curated reference model (e.g., model iJO1366[4]). To achieve a more comprehensive evaluation of the application's effectiveness, the comparison should include models of varying quality (highly curated and poorly curated) from organisms of different types (eukaryotes and prokaryotes, well-studied and less-studied). The criteria for comparing models may include, for instance, the number and presence or absence of metabolites, genes, and reactions [30, 27], the ability to directly perform FBA (Flux Balance Analysis) [28, 21], and reports from tools such as FROG [31] and MEMOTE [32], which provide additional insights into the metabolic network (e.g., mass and charge balancing, stoichiometric consistency, FVA, and gene/reaction deletion fluxes).

[3] *E. Coli* genome is available at NCBI website: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000005845.2

[4] This is the most recent *E. Coli* model available in BiGG: http://bigg.ucsd.edu/models/iJO1366

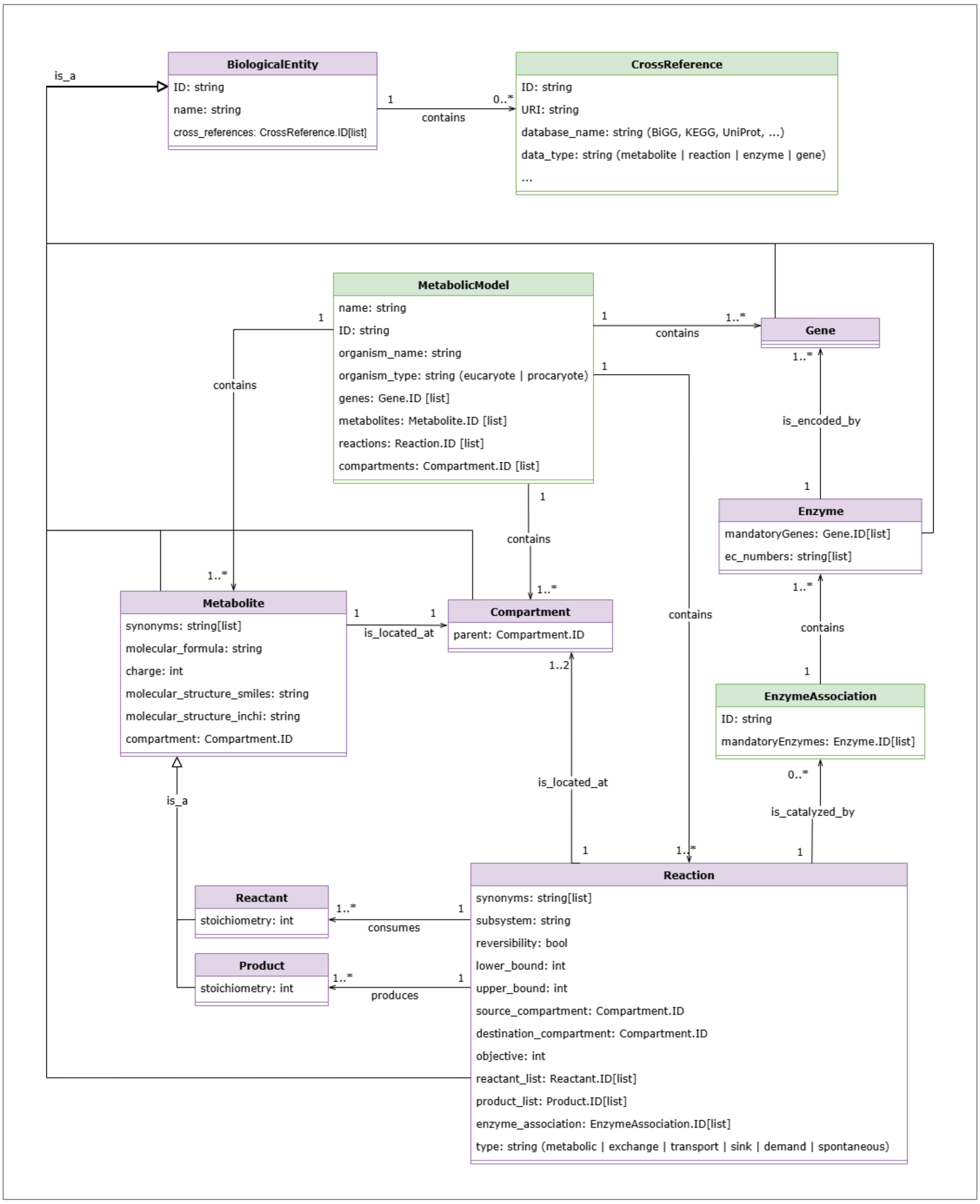

BiologicalEntity
ID: string
name: string
cross_references: CrossReference.ID[list]
is_a
1
contains
0..*
CrossReference
ID: string
URI: string
database_name: string (BiGG, KEGG, UniProt, ...)
data_type: string (metabolite | reaction | enzyme | gene)
...
MetabolicModel
name: string
ID: string
organism_name: string
organism_type: string (eucaryote | procaryote)
genes: Gene.ID [list]
metabolites: Metabolite.ID [list]
reactions: Reaction.ID [list]
compartments: Compartment.ID [list]
1
contains
1..*
Gene
1..*
is_encoded_by
1
Enzyme
mandatoryGenes: Gene.ID[list]
ec_numbers: string[list]
1..*
contains
1
EnzymeAssociation
ID: string
mandatoryEnzymes: Enzyme.ID[list]
0..*
is_catalyzed_by
1
contains
1..*
Metabolite
synonyms: string[list]
molecular_formula: string
charge: int
molecular_structure_smiles: string
molecular_structure_inchi: string
compartment: Compartment.ID
1
1
is_located_at
Compartment
parent: Compartment.ID
1..2
is_located_at
1
1..*
contains
Reaction
synonyms: string[list]
subsystem: string
reversibility: bool
lower_bound: int
upper_bound: int
source_compartment: Compartment.ID
destination_compartment: Compartment.ID
objective: int
reactant_list: Reactant.ID[list]
product_list: Product.ID[list]
enzyme_association: EnzymeAssociation.ID[list]
type: string (metabolic | exchange | transport | sink | demand | spontaneous)
is_a
Reactant
stoichiometry: int
1..*
consumes
1
Product
stoichiometry: int
1..*
produces
1

Figure 2: Initial domain model. Classes are represented as boxes; white-tipped arrows indicate inheritance, while standard arrows denote associations along with their cardinalities. Colors were used to distinguish between classes representing biological entities and those representing non-biological entities.

3.3. Application

To seamlessly integrate the GEM ontology into the GEM reconstruction workflow, a user-friendly software application must be developed in the future. This application should be able to load a draft model (any valid SBML file) and automatically retrieve additional data from biochemical databases (e.g., BiGG, KEGG, MetaCyc, MetaNetX) and relevant ontologies (e.g., Gene Ontology, ChEBI) to enrich the model and build a standardized representation, while the ontology would provide a formal structure for the data, facilitating data integration and logical consistency verification. The application should also generate a report summarizing reconciliation results – logging automated decisions, describing unresolved issues, providing qualitative and quantitative model evaluations, and presenting additional information to assist domain experts in refining the model, either manually or automatically, depending on the nature of the issues detected. Finally, it should also allow the experts to edit, refine, and subsequently export the model in multiple formats.

The technical details of such an application should be defined during the implementation phase. This includes selecting the most appropriate biochemical databases and ontologies for data enrichment, defining mechanisms for data access (e.g., APIs, SPARQL endpoints), and establishing strategies for data management and storage. In addition, the deployment and documentation of the application should be carefully planned to ensure long-term maintainability, reproducibility, and ease of use for both users and developers.

4. Conclusion

This work proposed an ontology-based approach to streamline the reconstruction of genome-scale metabolic models (GEMs), comprising both a GEM ontology and an application to integrate it into the reconstruction workflow. The ontology – currently under development – is expected to facilitate data reconciliation across biochemical databases and enable automated reasoning. The application is intended to centralize information retrieval from multiple datasets, thereby reducing manual effort and improving efficiency. Developing both the ontology and the application is complex and time-consuming; therefore, the work can be divided into two phases. The first phase should focus on ontology development and evaluation, while the second should focus on building an application on top of the ontology to generate enhanced models.

In addition, the comparison of GEMs remains an open challenge, as it requires the unambiguous identification of model components such as metabolites and reactions. Defining comparison criteria for metabolic models is expected to provide a novel standard for evaluating the quality of GEMs generated by different tools in future studies.

Acknowledgments

This work was supported by a doctoral scholarship from CAPES/PROEX and CNPq (140323/2025-2), and by funding from Inria ERABLE – University of Lyon, which enabled a research visit to the LBBE (Laboratoire de Biométrie et Biologie Évolutive) in Lyon, France, to foster collaboration between our research teams.

Special thanks to Gabriela Torres Montanaro (ICB-USP), Ariel Mariano Silber (ICB-USP), Lucas Gentil Azevedo (Cidacs/FIOCRUZ, BA), Mariana Galvão Ferrarini (Max Planck Institute for Chemical Ecology in Jena, Germany) and Marie-France Sagot (University of Lyon, France) for their valuable contributions to this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (GPT-4o) and Claude Sonnet 4 in order to: Grammar and spelling check, paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Judge, M. S. Dodd, Metabolism, Essays in Biochemistry 64 (2020) 607–647. URL: https://doi.org/10.1042/EBC20190041. doi:10.1042/EBC20190041.

[2] T. Bernard, A. Bridge, A. Morgat, S. Moretti, I. Xenarios, M. Pagni, Reconciliation of metabolites and biochemical reactions for metabolic networks, Briefings in Bioinformatics 15 (2012) 123–135. URL: https://doi.org/10.1093/bib/bbs058. doi:10.1093/bib/bbs058.

[3] A. Kumar, P. F. Suthers, C. D. Maranas, Metrxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases, BMC Bioinformatics 13 (2012) 6. URL: https://doi.org/10.1186/1471-2105-13-6. doi:10.1186/1471-2105-13-6.

[4] A. Hari, A. Zarrabi, D. Lobo, mergem: merging, comparing, and translating genome-scale metabolic models using universal identifiers, NAR Genomics and Bioinformatics 6 (2024) 1–21. URL: https://doi.org/10.1093/nargab/lqae010. doi:10.1093/nargab/lqae010.

[5] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: a database and ontology for chemical entities of biological interest, Nucleic Acids Research 36 (2007) D344–D350. URL: https://doi.org/10.1093/nar/gkm791. doi:10.1093/nar/gkm791.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. URL: https://doi.org/10.1038/75556. doi:10.1038/75556.

[7] M. Courtot, N. Juty, C. Knüpfer, D. Waltemath, A. Zhukova, A. Dräger, M. Dumontier, A. Finney, M. Golebiewski, J. Hastings, S. Hoops, S. Keating, D. B. Kell, S. Kerrien, J. Lawson, A. Lister, J. Lu, R. Machne, P. Mendes, M. Pocock, N. Rodriguez, A. Villeger, D. J. Wilkinson, S. Wimalaratne, C. Laibe, M. Hucka, N. Le Novère, Controlled vocabularies and semantics in systems biology, Molecular Systems Biology 7 (2011) 543. URL: https://www.embopress.org/doi/abs/10.1038/msb.2011.77. doi:https://doi.org/10.1038/msb.2011.77.

[8] I. Thiele, B. Ø. Palsson, A protocol for generating a high-quality genome-scale metabolic reconstruction, Nature Protocols 5 (2010) 93–121. URL: https://doi.org/10.1038/nprot.2009.203. doi:10.1038/nprot.2009.203.

[9] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, Journal of Chemical Information and Computer Sciences 28 (1988) 31–36. URL: https://doi.org/10.1021/ci00057a005. doi:10.1021/ci00057a005.

[10] A. McNaught, The iupac international chemical identifier: Inchl-a new standard for molecular informatics, Chemistry international 28 (2006) 12–14. URL: https://api.semanticscholar.org/CorpusID:89057853.

[11] L. Cottret, P. Vieira Milreu, V. Acuña, A. Marchetti-Spaccamela, F. Viduani Martinez, M.-F. Sagot, L. Stougie, Enumerating precursor sets of target metabolites in a metabolic network, in: K. A. Crandall, J. Lagergren (Eds.), Algorithms in Bioinformatics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 233–244.

[12] A. Amara, C. Frainay, F. Jourdan, T. Naake, S. Neumann, E. M. Novoa-del Toro, R. M. Salek, L. Salzer, S. Scharfenberg, M. Witting, Networks and graphs discovery in metabolomics data analysis and interpretation, Frontiers in Molecular Biosciences 9 (2022). URL: https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2022.841373. doi:10.3389/fmolb.2022.841373.

[13] V. Lacroix, L. Cottret, P. Thébault, M.-F. Sagot, An introduction to metabolic networks and their structural analysis, IEEE/ACM Transactions on Computational Biology and Bioinformatics 5 (2008) 594–617. doi:10.1109/TCBB.2008.79.

[14] L. Wang, S. Dash, C. Y. Ng, C. D. Maranas, A review of computational tools for design and reconstruction of metabolic pathways, Synthetic and Systems Biotechnology 2 (2017) 243–252. URL: https://www.sciencedirect.com/science/article/pii/S2405805X17300820. doi:https://doi.org/10.1016/j.synbio.2017.11.002.

[15] L. Strömbäck, P. Lambrix, Representations of molecular pathways: an evaluation of sbml, psi mi and biopax, Bioinformatics 21 (2005) 4401–4407. URL: https://doi.org/10.1093/bioinformatics/bti718. doi:10.1093/bioinformatics/bti718.

[16] T. Pfau, M. P. Pacheco, T. Sauter, Towards improved genome-scale metabolic network recon-

structions: unification, transcript specificity and beyond, Briefings in Bioinformatics 17 (2015) 1060–1069. URL: https://doi.org/10.1093/bib/bbv100. doi:10.1093/bib/bbv100.

[17] M. A. Carey, A. Dräger, M. E. Beber, J. A. Papin, J. T. Yurkovich, Community standards to facilitate development and address challenges in metabolic modeling, Molecular Systems Biology 16 (2020) e9235. URL: https://www.embopress.org/doi/abs/10.15252/msb.20199235. doi:https://doi.org/10.15252/msb.20199235.

[18] D. B. Bernstein, S. Sulheim, E. Almaas, D. Segrè, Addressing uncertainty in genome-scale metabolic model reconstruction and analysis, Genome Biology 22 (2021) 64. URL: https://doi.org/10.1186/s13059-021-02289-z. doi:10.1186/s13059-021-02289-z.

[19] N. Pham, R. G. A. van Heck, J. C. J. van Dam, P. J. Schaap, E. Saccenti, M. Suarez-Diez, Consistency, Inconsistency, and Ambiguity of Metabolite Names in Biochemical Databases Used for Genome-Scale Metabolic Modelling, Metabolites 9 (2019). URL: https://www.mdpi.com/2218-1989/9/2/28. doi:10.3390/metabo9020028.

[20] G. Guizzardi, N. Guarino, Explanation, semantics, and ontology, Data and Knowledge Engineering 153 (2024) 102325. URL: https://www.sciencedirect.com/science/article/pii/S0169023X24000491. doi:https://doi.org/10.1016/j.datak.2024.102325.

[21] S. M. D. Seaver, F. Liu, Q. Zhang, J. Jeffryes, J. P. Faria, J. N. Edirisinghe, M. Mundy, N. Chia, E. Noor, M. E. Beber, A. A. Best, M. DeJongh, J. A. Kimbrel, P. D'haeseleer, S. R. McCorkle, J. R. Bolton, E. Pearson, S. Canon, E. M. Wood-Charlson, R. W. Cottingham, A. P. Arkin, C. S. Henry, The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes, Nucleic Acids Research 49 (2020) D575–D588. URL: https://doi.org/10.1093/nar/gkaa746. doi:10.1093/nar/gkaa746.

[22] J. N. Otte, J. Beverley, A. Ruttenberg, Bfo: Basic formal ontology, Applied ontology 17 (2022) 17–43. doi:10.3233/ao-220262.

[23] M. Hucka, F. T. Bergmann, C. Chaouiya, A. Dräger, S. Hoops, S. M. Keating, M. König, N. L. Novère, C. J. Myers, B. G. Olivier, S. Sahle, J. C. Schaff, R. Sheriff, L. P. Smith, D. Waltemath, D. J. Wilkinson, F. Zhang, The systems biology markup language (sbml): Language specification for level 3 version 2 core release 2, Journal of Integrative Bioinformatics 16 (2019) 20190021. URL: https://doi.org/10.1515/jib-2019-0021. doi:doi:10.1515/jib-2019-0021.

[24] A. García S., A. Bernasconi, G. Guizzardi, O. Pastor, V. C. Storey, I. Panach, Assessing the value of ontologically unpacking a conceptual model for human genomics, Information Systems 118 (2023) 102242. URL: https://www.sciencedirect.com/science/article/pii/S0306437923000789. doi:https://doi.org/10.1016/j.is.2023.102242.

[25] A. Bernasconi, G. Guizzardi, O. Pastor, V. C. Storey, Semantic interoperability: ontological unpacking of a viral conceptual model, BMC Bioinformatics 23 (2022) 491. URL: https://doi.org/10.1186/s12859-022-05022-0. doi:10.1186/s12859-022-05022-0.

[26] G. K. Q. Monfardini, J. S. Salamon, M. P. Barcellos, Use of competency questions in ontology engineering: A survey, in: J. P. A. Almeida, J. Borbinha, G. Guizzardi, S. Link, J. Zdravkovic (Eds.), Conceptual Modeling, Springer Nature Switzerland, Cham, 2023, pp. 45–64.

[27] S. N. Mendoza, B. G. Olivier, D. Molenaar, B. Teusink, A systematic assessment of current genome-scale metabolic reconstruction tools, Genome Biology 20 (2019) 158. URL: https://doi.org/10.1186/s13059-019-1769-1. doi:10.1186/s13059-019-1769-1.

[28] J. Zimmermann, C. Kaleta, S. Waschina, gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models, Genome Biology 22 (2021) 81. URL: https://doi.org/10.1186/s13059-021-02295-1. doi:10.1186/s13059-021-02295-1.

[29] M. Di Filippo, C. Damiani, D. Pescini, Gpruler: Metabolic gene-protein-reaction rules automatic reconstruction, PLOS Computational Biology 17 (2021) 1–25. URL: https://doi.org/10.1371/journal.pcbi.1009550. doi:10.1371/journal.pcbi.1009550.

[30] Y. E. Hsieh, K. Tandon, H. Verbruggen, Z. Nikoloski, Comparative analysis of metabolic models of microbial communities reconstructed from automated tools and consensus approaches, npj Systems Biology and Applications 10 (2024) 54. URL: https://doi.org/10.1038/s41540-024-00384-y.

doi:10.1038/s41540-024-00384-y.

[31] K. Raman, M. Kratochvíl, B. G. Olivier, M. König, P. Sengupta, D. K. Kuppa Baskaran, T. V. N. Nguyen, D. Lobo, S. E. Wilken, K. K. Tiwari, A. K. Raghu, I. Palanikumar, L. Raajaraam, M. Ibrahim, S. Balakrishnan, S. Umale, F. Bergmann, T. Malpani, V. P. Satagopam, R. Schneider, M. E. Beber, S. Keating, M. Anton, A. Renz, M. Lakshmanan, D.-Y. Lee, L. Koduru, R. Mostolizadeh, O. Dias, E. Cunha, A. Oliveira, Y. Q. Lee, K. Zengler, R. Santibáñez-Palominos, M. Kumar, M. Barberis, B. L. Puniya, T. Helikar, H. V. Dinh, P. F. Suthers, C. D. Maranas, I. Casini, S. B. Loghmani, N. Veith, N. Leonidou, F. Li, Y. Chen, J. Nielsen, G. Lee, S. M. Lee, G. B. Kim, P. T. Monteiro, M. C. Teixeira, H. U. Kim, S. Y. Lee, U. W. Liebal, L. M. Blank, C. Lieven, C. Tarzi, C. Angione, M. E. Blaise, Ç. P. Aytar, M. Kulyashov, l. Akberdin, D. Kim, S. H. Yoon, Z. Xu, J. Gautam, W. T. Scott, P. J. Schaap, J. J. Koehorst, C. Zuñiga, G. Canto-Encalada, S. Benito-Vaquerizo, I. P. Olm, M. Suarez-Diez, Q. Yuan, H. Ma, M. M. Islam, J. A. Papin, F. Zorrilla, K. R. Patil, A. Basile, J. Nogales, G. S. León, F. Castillo-Alfonso, R. Olivares-Hernández, G. Canto-Encalada, G. Vigueras-Ramírez, H. Hermjakob, A. Dräger, R. S. Malik-Sheriff, Frog analysis ensures the reproducibility of genome scale metabolic models, bioRxiv (2024). URL: https://www.biorxiv.org/content/early/2024/09/26/2024.09.24.614797. doi:10.1101/2024.09.24.614797.

[32] C. Lieven, M. Beber, B. Olivier, F. Bergmann, M. Ataman, P. Babaei, J. Bartell, L. Blank, S. Chauhan, K. Correia, C. Diener, A. Dräger, B. Ebert, J. Edirisinghe, J. Faria, A. Feist, G. Fengos, R. Fleming, B. García-Jiménez, V. Hatzimanikatis, W. van Helvoirt, C. Henry, H. Hermjakob, M. Herrgård, A. Kaafarani, H. Kim, Z. King, S. Klamt, E. Klipp, J. Koehorst, M. König, M. Lakshmanan, D.-Y. Lee, S. Lee, S. Lee, N. Lewis, F. Liu, H. Ma, D. Machado, R. Mahadevan, P. Maia, A. Mardinoglu, G. Medlock, J. Monk, J. Nielsen, L. Nielsen, J. Nogales, I. Nookaew, B. Palsson, J. Papin, K. Patil, M. Poolman, N. Price, O. Resendis-Antonio, A. Richelle, I. Rocha, B. Sánchez, P. Schaap, R. Malik Sheriff, S. Shoaie, N. Sonnenschein, B. Teusink, P. Vilaça, J. Vik, J. Wodke, J. Xavier, Q. Yuan, M. Zakhartsev, C. Zhang, Memote for standardized genome-scale metabolic model testing, Nature Biotechnology 38 (2020). doi:10.1038/s41587-020-0446-y.

OntoMI: An ontology grounded in the theory of multiple intelligences for semantic classification of educational resources

Jefferson Rodrigo Speck[1,*], Sidgley Camargo de Andrade[2] and Clodis Boscarioli[1]

[1] Western Paraná State University (Unioeste), Master's Program in Computer Science, P.O. Box 711, 85819-110, Cascavel – PR, Brazil
[2] Federal University of Technology - Parana (UTFPR), Toledo Campus, 19 Cristo Rei Street, Vila Becker, 85902-490, Toledo – PR, Brazil

Abstract

This article introduces OntoMI, a semantic ontology based on Howard Gardner's Theory of Multiple Intelligences, developed to formally represent and infer the cognitive dimensions evoked by educational texts. OntoMI provides an organized conceptual framework that enables the identification, classification and quantification of multiple intelligences in educational texts. It serves as the basis for a computerized model that processes texts, extracts elements and infers cognitive activations through semantic inferences. Based on these inferences, the system creates explainable cognitive profiles for each resource, which are represented as intelligence distribution vectors. This approach aims to enable the semantic classification and evaluation of content to support more comprehensive pedagogical analysis, personalized access to learning materials and adaptation to individual cognitive profiles.

Keywords

Ontology, Multiple Intelligences, Educational Technology, Semantic Classification, Personalized Learning

1. Introdução

The uniqueness of human beings manifests itself in several dimensions — cognitive, affective, social and cultural — that have a direct impact on how individuals learn and interact with knowledge [1, 2, 3]. This diversity requires pedagogical approaches that not only recognize these differences, but operationalize them as central elements in the planning and delivery of instruction. The Theory of Multiple Intelligences (MI) proposed by Howard Gardner offers a conceptual framework for this, which assumes that human cognition manifests itself in different areas of competence, such as linguistic, logical-mathematical, musical, spatial, physical-kinesthetic, interpersonal, intrapersonal, naturalistic and existential intelligences [3].

Despite the growing demand for personalized education systems, most current approaches still rely on standardized teaching models that ignore the diversity of individuals' learning styles and processes. Even when some degree of customization is attempted, the appropriation of theory is usually limited and superficial, which restricts its application. In digital contexts, this limitation is exacerbated by the lack of models capable of representing, deriving and applying the principles of MI to the analysis or recommendation of instructional content in a structured, explainable and scalable way. This gap hinders the advancement of pedagogical practices that respond to cognitive plurality in light of MI, and complicates the identification, classification and use of materials based on the intelligences they elicit — especially on a large scale and with computerized support.

Against this background, the present work proposes the development of OntoMI, a semantic ontology based on the MI theory and aimed at the formal representation of the cognitive dimensions elicited by educational texts. The proposal addresses the following central research question: *How can textual*

Proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and 9th Doctoral and Masters Consortium on Ontologies (WTDO 2025), São José dos Campos (SP), Brazil, September 29 – October 02, 2025.

[*] Corresponding author.

jeffersonspeck@msn.com (J. R. Speck); sidgleyandrade@utfpr.edu.br (S. C. d. Andrade); clodis.boscarioli@unioeste.br (C. Boscarioli)
0009-0009-3964-9185 (J. R. Speck); 0000-0003-0371-7835 (S. C. d. Andrade); 0000-0002-7110-2026 (C. Boscarioli)

educational resources be semantically classified to support personalized teaching while remaining faithful to the Theory of Multiple Intelligences?

OntoMI attempts to fill this gap by providing an ontological infrastructure that enables the identification, classification and quantification of features of intelligences elicited through the semantic mapping of textual elements to ontological classes and properties. Its construction is based on the systematization of the pedagogical principles contained in Gardner's works and on a conceptual modeling oriented towards inference that allows educational materials to be interpreted according to the dominance of certain intelligences.

Therefore, this study aims to develop a formal semantic ontology based on Gardner's theory that is capable of representing, inferring and quantifying the cognitive dimensions evoked by educational textual content and that can be integrated into a computerized system. The specific aims of this study are: (OE1) to identify and systematize the pedagogical foundations of MI directly from Gardner's works; (OE2) to develop a semantic ontology that focuses on the representation of MI in educational contexts; and (OE3) to implement a computational model for classifying educational texts based on OntoMI.

The aim is to provide a conceptual and technical tool capable of matching educational content with students' cognitive profiles and supporting pedagogical curation and personalized teaching from an explainable, semantically structured perspective coherent with the principles of the theory.

2. Related work

Several studies have used ontologies as the basis for adaptive educational systems and have investigated their ability to formally represent knowledge and allow conclusions to be drawn about content and learning profiles. One example is the work of Vasiliki Demertzi and Konstantinos Demertzis [4], who propose an adaptive teaching system based on ontological matching that enables personalized content recommendations according to the mapping between students and teaching materials. Although it contributes to personalized learning, their proposal takes an ontology-centric approach that focuses on the scope of the study and does not involve cognitive concepts.

Similarly, Monika Rani et al. [5] presented the OPAESFH system, which combines ontologies with inference techniques based on Fuzzy Petri Nets (FPN) and Hidden Markov Models (HMM) to adapt instruction to student characteristics. Despite its technical sophistication, the model does not integrate a conceptual structure based on cognitive theories and is limited to predefined learning profile categories.

The work of Pornpit Wongthongtham et al. [6] explicitly attempts to integrate MI theory into a fuzzy ontology aimed at the semantic annotation of educational content. The proposal highlights the potential of MI as a basis for intelligent recommender systems and personalized learning, but still lacks a structured ontological formalization of intelligences, especially one aimed at the detailed analysis of textual materials.

Against this background, this paper proposes OntoMI, a semantic ontology developed based on the original principles of the MI theory proposed by Howard Gardner. In contrast to the aforementioned approaches, OntoMI aims to formally represent and infer the MI evoked by textual educational content by providing an explainable conceptual and computational infrastructure capable of generating cognitive vectors expressing the distribution of intelligences activated by each resource. As such, the proposal brings advances in terms of theoretical fidelity to MI, semantic classification capability, and the development of educational systems more attuned to cognitive diversity.

3. Methodology

The methodological approach of this applied research is aimed at solving a practical problem related to personalized learning and the pedagogical curation of learning objects based on the MI theory. The investigation is structured in complementary phases that include a conceptual foundation, a review of the state of the art, and the development of computational artifacts. A summary of this methodological structure is presented in Figure 1.

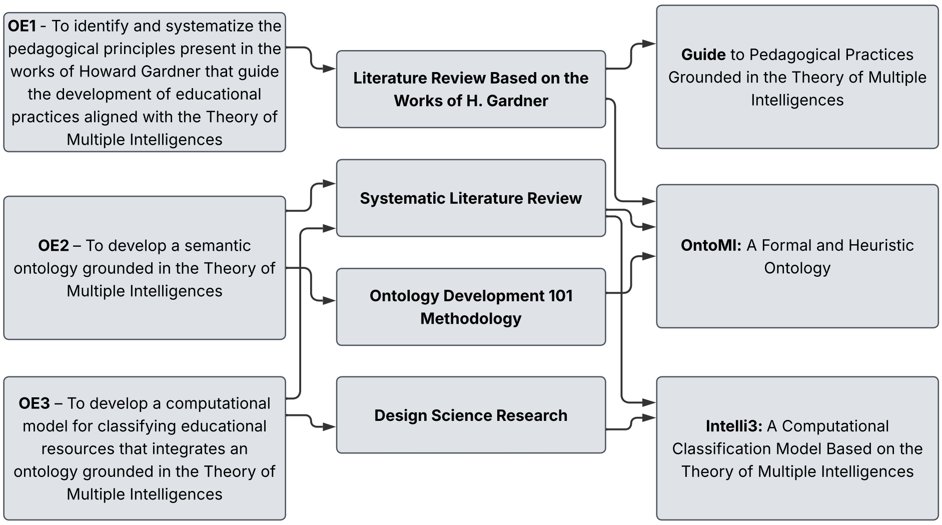

OE1 - To identify and systematize the pedagogical principles present in the works of Howard Gardner that guide the development of educational practices aligned with the Theory of Multiple Intelligences
OE2 – To develop a semantic ontology grounded in the Theory of Multiple Intelligences
OE3 – To develop a computational model for classifying educational resources that integrates an ontology grounded in the Theory of Multiple Intelligences
Literature Review Based on the Works of H. Gardner
Systematic Literature Review
Ontology Development 101 Methodology
Design Science Research
Guide to Pedagogical Practices Grounded in the Theory of Multiple Intelligences
OntoMI: A Formal and Heuristic Ontology
Intelli3: A Computational Classification Model Based on the Theory of Multiple Intelligences

Figure 1: Methodological model of research: objectives, methods and artifacts. Source: authors' elaboration.

The starting point was an in-depth analysis of the works of Howard Gardner, focusing on the epistemological and pedagogical assumptions of MI theory. This formed the basis not only for the systematization of the pedagogical criteria (OE1), but also for the conceptual support necessary for structuring the ontology (OE2). This first phase is characterized by an exploratory approach aimed at directly extracting the core elements of the theory in relation to pedagogical practices, avoiding secondary interpretations or purely instrumental uses. As a result of this phase, a guide to pedagogical practices based on MI has been developed, containing principles and guiding criteria for planning teaching and learning experiences adapted to the cognitive diversity of learners.

Subsequently, a systematic literature review (SLR) was conducted, following the methodological guidelines of Barbara A. Kitchenham, David Budgen, and Pearl Brereton [7], with the aim of critically capturing how MI is used in digital educational environments. The SLR identified recurring gaps in the computational application of theory, particularly in relation to the lack of semantic mechanisms, limited personalization strategies, and a lack of structures capable of deriving cognitive profiles from textual data. This mapping supported the conceptual and technical choices underlying the proposals described in OE2 and OE3.

On this basis, pedagogical and computational artifacts are being developed using methods appropriate to the nature of each phase of the research. In line with OE2, OntoMI has been developed — a formal and heuristic ontology created according to the *Ontology Development 101* methodology [8] and adapted for the semantic representation of text elements that evoke different intelligences. This ontology represents the main conceptual artifact of the research and enables the modeling of inferential relationships between theoretical concepts and the construction of cognitive vectors describing the profiles activated by educational content. It is validated by analyzing illustrative examples, checking semantic coherence, conceptual coverage, and computational applicability.

Finally, in response to OE3, the research culminates in the development of the computational model *Intelli3*, built using the *Design Science Research* (DSR) methodology [9], which focuses on the construction and evaluation of technological artifacts to solve practical problems. The system represents the main computational artifact of the study and is structured by a modular, multi-layered architecture that ensures flexibility, scalability and separation of functional responsibilities. This architecture was designed to operationalize MI at a computational scale.

A focus group composed of educators and MI specialists will be formed to validate the artifacts of the study and the data obtained during testing. The evaluation will follow a knowledge elicitation methodology based on the Delphi [10, 11] qualified consensus on the conceptual coherence, pedagogical applicability and appropriateness of the conclusions drawn by the system.

4. The OntoMI ontology

OntoMI is a formal and heuristic semantic ontology that was developed to conceptualize and infer in a structured way the MI elicited by educational content expressed in natural language. The ontology is based on the principles of MI theory, as proposed by Howard Gardner [2?], and aims to translate human cognitive diversity into an ontological architecture capable of supporting explainable mechanisms for analyzing and classifying textual educational resources.

The construction of OntoMI followed the principles of the *Ontology Development 101* methodology [8], which was adapted to the educational domain with a focus on the semantic representation of cognitive properties. This methodology was selected because of its simplicity and step-by-step orientation, which makes it particularly suitable for the creation of initial ontological artifacts and for maintaining clarity in scope definition. The process comprised: (i) a clear specification of the domain and goals of the ontology; (ii) the identification and organization of recurring terms and concepts in pedagogical discourse; (iii) the definition of semantic categories related to the intelligences proposed by Gardner; and (iv) the modeling of classes, properties, and axioms that enable the derivation of cognitive profiles from observable linguistic elements.

The conceptual structure of OntoMI is organized around three main types of elements extracted from texts. The first are `keywords`, which correspond to terms that represent concepts, content, or cognitive operations strongly associated with specific intelligences. The second are `ContextObjects`, which denote the central topics of the content and their disciplinary connections. Finally, there are the `DiscursiveStrategies`, referring to the ways in which the content is organized and presented, such as through narratives, descriptions, or comparisons.

Each of these elements, once identified in a text segment, is linked to one or more intelligences via the `evokesIntelligence` property. This relationship is not binary, but weighted: Each association can have a certain weight that reflects the intensity with which the element evokes a certain intelligence. The exact definition of this weighting is left to the person performing the inference, which allows for flexibility in the application of the model. However, for the purposes of this study, the weights are discussed and determined with a focus group.

Based on the co-occurrence and intensity of the elements, the OntoMI computational system generates instances of the class `IntelligenceActivation`, which formalizes the inference that a given text fragment cognitively activates one or more intelligences. Figure 2 depicts the conceptual model of OntoMI in UML and highlights its main classes and ontological relationships.

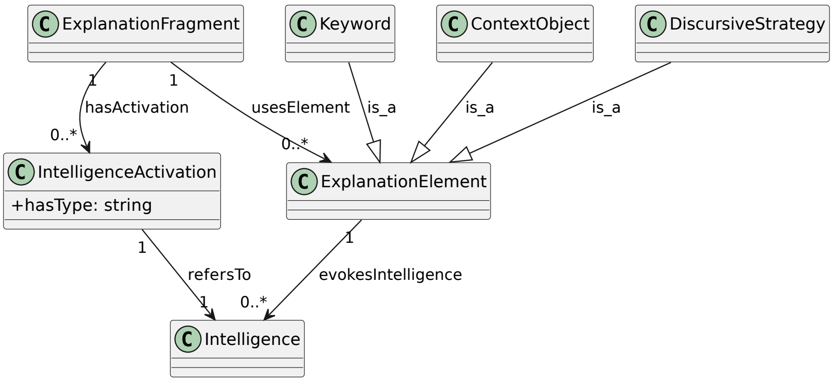


Figure 2: Conceptual model of OntoMI: UML structure of the inferential relationships between text fragments and multiple intelligences.

To summarize, a natural language educational content is broken down into explanatory fragments during processing, from which three main types of elements are identified: Keywords, Central Themes and Discursive Strategies. Each of these elements is assigned to a corresponding ontology class — `Keyword, ContextObject or DiscursiveStrategy`— associated with one or more of the MI proposed by Gardner via the property `evokesIntelligence`. This association is weighted by heuristic values that indicate the intensity of the cognitive stimulation. The result of this process is the creation of instances of the class `IntelligenceActivation`, which formally and explainably represent which intelligences are activated by a particular text segment. This mechanism aims to convert content into interpretable semantic representations that can be used by computational systems focused on analysis.

OntoMI was developed to be integrated into computer systems for text analysis in education, such as the cognitive classifier *Intelli3* proposed here. The central function of ontology in this context is to provide a formal basis for systems to identify, classify and semantically quantify intelligences elicited by textual educational content. The integration between the ontology and computer models should enable the generation of explainable cognitive vectors — vector representations of the distribution of intelligences in a given resource that can be used in various pedagogical applications such as personalized instruction, curriculum analysis, and semantic indexing of learning objects.

To illustrate how OntoMI is instantiated in practice, consider the fragment "Clapping hands, students imitated the constant motion of a body without external interference." In this example, the linguistic elements are mapped to `Keyword` (e.g., *motion*, *body*), a `ContextObject` (*Newton's first law – inertia*), and a `DiscursiveStrategy` (*imitation/experiential activity*). These instances feed an `IntelligenceActivation`, which—via weighted links—evokes bodily-kinesthetic and logical-mathematical intelligences. The figure below summarizes this minimal instantiation and the corresponding inference flow.

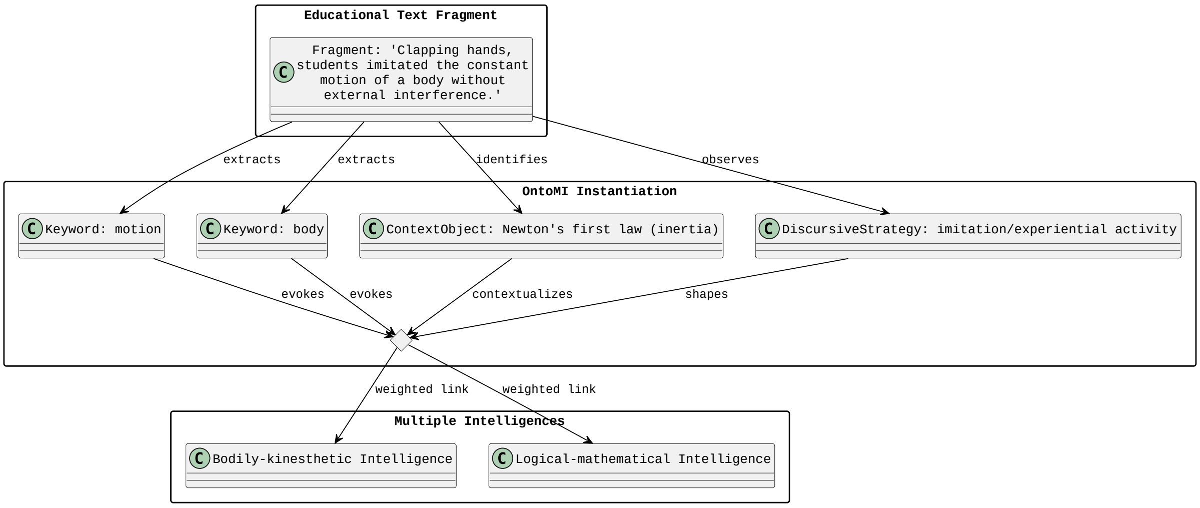

Educational Text Fragment
Fragment: 'Clapping hands, students imitated the constant motion of a body without external interference.'
extracts
extracts
identifies
observes
OntoMI Instantiation
Keyword: motion
Keyword: body
ContextObject: Newton's first law (inertia)
DiscursiveStrategy: imitation/experiential activity
evokes
evokes
contextualizes
shapes
weighted link
weighted link
Multiple Intelligences
Bodily-kinesthetic Intelligence
Logical-mathematical Intelligence

Figure 3: Minimal OntoMI instantiation.

4.1. Data collection

Data collection for the application and validation of OntoMI is carried out through the selection of educational materials in text form, which include textbooks, handouts, scientific articles and lesson plans from various subject areas (preferably in editable formats such as PDF, TXT, or HTML), transcripts of video lectures and other discursive resources available in public repositories, as well as learning objects and materials from freely accessible educational platforms, always considering usage licenses and public domain availability. These materials are organized, segmented, and, when necessary, manually annotated to ensure quality in the application of the ontology and in the creation of cognitive vectors.

4.2. Ontology validation

The validation of OntoMI will be carried out in two complementary stages. The first consists of verifying the structural and semantic consistency of the ontology independently, following the criteria of *Ontology Development 101*, ensuring clarity of scope, coherence of relations, completeness, and absence of ambiguities. The second stage involves applying OntoMI within the computational classifier, initially using large language models (LLM) to support the instantiation of textual fragments. This phase will be evaluated through proof-of-concept experiments and expert analysis in a focus group. In both stages, the evaluation will consider semantic adherence, conceptual coverage, explainability, and pedagogical applicability, verifying whether the ontology adequately represents multiple intelligences and supports reliable inference over educational texts.

Computational Model Validation: While the focus is on the ontology, the computational validation phase helps to demonstrate the practical applicability of OntoMI and its potential as a basis for educational systems that take cognitive diversity into account.

5. Considerations and next steps

The research has already made significant progress towards its specific objectives. OE1, which involved identifying and systematizing the pedagogical principles in Howard Gardner's works, has been fully achieved. As a product of this phase, the Guide to Pedagogical Practices based on MI was developed, which summarizes criteria and strategies related to cognitive diversity and serves as a theoretical validation basis for the other project artifacts.

OE2, which focuses on the development of a semantic ontology based on MI, is at an advanced stage of implementation. The conceptual structure of OntoMI has been defined, including the modeling of key ontological classes, semantic properties, inference rules and heuristic weightings related to the activation of intelligence. Work is currently underway on the integration of inference elements into the formal OWL structure and preparations for practical use in the classification system (Figure 4).

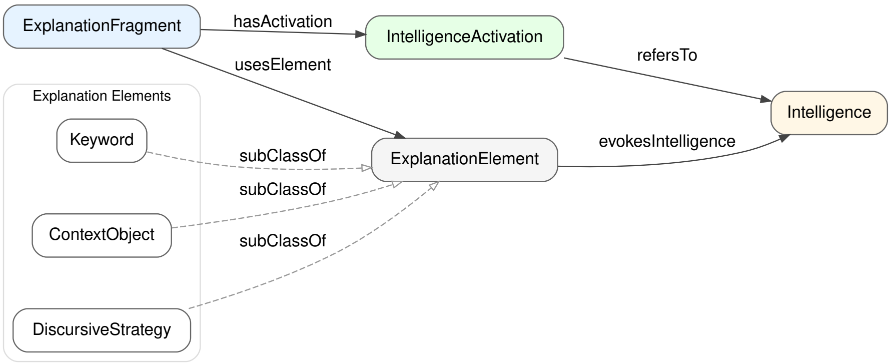


Figure 4: OntoMI model: OWL structure of the inferential relationships between text fragments and multiple intelligences.

In parallel, OE3 — which proposes the development of a computational model for classifying educational resources based on OntoMI — has already defined its architecture, summarized in the following pipeline: input educational text → semantic segmentation → OntoMI instantiation → inference → cognitive vector. The *Intelli3* system parses textual resources, segments them into explanatory fragments, maps linguistic elements (keywords, context objects, discursive strategies) to OntoMI classes, and applies inference rules that generate `IntelligenceActivation` instances. These are aggregated into cognitive vectors that represent the distribution of intelligences across the text and enable similarity measures and personalized recommendations aligned with students' profiles.

The integration between OE2 and OE3 is already planned, and the next steps of the research will focus on completing the operational ontology, developing semantic inference mechanisms, and functionally validating the *Intelli3* system with real educational materials. Still pending are the structured collection of natural language educational data and the selection of experts for the focus group, who will participate in the qualitative evaluation of both the ontology and the system results, including the weighting analysis of the guide. These activities will be carried out in parallel with testing and proof-of-concept validation.

Declaration on Generative AI

During the preparation of this work, the author(s) used artificial intelligence tools to assist with translation and language editing. Specifically, ChatGPT was utilized to translate the manuscript into English, and InstaText.io was used to refine the grammar and phrasing. The author(s) reviewed and edited the final output and take(s) full responsibility for the content of the publication.

References

[1] H. Gardner, Frames of mind: The theory of multiple intelligences, 2nd ed., Basic Books, New York, 2011.

[2] H. Gardner, Multiple intelligences: New horizons, Basic Books, New York, USA, 2006.

[3] H. Gardner, Intelligence reframed: Multiple intelligences for the 21st century, Basic Books, 2000.

[4] V. Demertzi, K. Demertzis, A hybrid ontology matching mechanism for adaptive educational elearning environments, International Journal of Information Technology & Decision Making 22 (2023) 1813–1841. doi:10.1142/S0219622022500936.

[5] M. Rani, R. Vyas, O. P. Vyas, Opaesfh: Ontology-based personalized adaptive e-learning system using fpn and hmm, in: TENCON 2017 - 2017 IEEE Region 10 Conference, 2017, pp. 2441–2446. doi:10.1109/TENCON.2017.8228271.

[6] P. Wongthongtham, K. Y. Chan, V. Potdar, B. Abu-Salih, S. Gaikwad, P. Jain, State-of-the-art ontology annotation for personalised teaching and learning and prospects for smart learning recommender based on multiple intelligence and fuzzy ontology, International Journal of Fuzzy Systems 20 (2018) 1357–1372. URL: https://doi.org/10.1007/s40815-018-0467-6. doi:10.1007/s40815-018-0467-6.

[7] B. Kitchenham, Guidelines for performing systematic literature reviews in software engineering, Technical Report EBSE-2007-01, EBSE Technical Report, Keele University and University of Durham, UK, 2007.

[8] N. F. Noy, D. L. McGuinness, Ontology development 101: A guide to creating your first ontology, Technical Report KSL-01-05, Stanford Knowledge Systems Laboratory, Stanford, CA, USA, 2001. Available at https://protege.stanford.edu/publications/ontology_development/ontology101.pdf.

[9] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Quarterly 28 (2004) 75–105. doi:10.2307/25148625.

[10] O. Helmer, Analysis of the future: The delphi method, The RAND Corporation, P-3558 (1967). A paper discussing the method's application and principles, building on earlier RAND work.

[11] H. A. Linstone, M. Turoff (Eds.), The delphi method: Techniques and applications, Addison-Wesley, 1975.

ONTOBRAS
SEMINAR ON ONTOLOGY RESEARCH IN BRAZIL

This volume contains the proceedings of the 18th Seminar on Ontology Research in Brazil (ONTOBRAS 2025) and the 9th Workshop on Theses and Dissertations in Ontologies (WTDO 2025), held in São José dos Campos (SP), Brazil, from September 29 to October 2, 2025.

The proceedings are published online in the CEUR Workshop Proceedings series (ISSN 1613-0073) and are available under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

CEUR Workshop Proceedings
ceur-ws.org
ISSN 1613-0073

I A O A
The International Association for Ontology and its Applications