Joint Conference
ICBO - ONTOBRAS
2023

Proceedings of the 14th International Conference on Biomedical Ontologies (ICBO 2023)

joint with the

Workshop on Ontologies for Infectious and Immune-Mediated Disease Data Science (OIIDDS 2023)

and the

FAIR Ontology Harmonization and Trust Data Interoperability Workshop (FOHTI 2023)

Hybrid conference, hosted in Brasília, Brazil and online

August 28 - September 1st, 2023

edited by
Fernanda Farinelli, Amanda Damasceno de Souza and Eduardo Ribeiro Felipe

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

Imprint

Publishers and Editors:

Fernanda Farinelli, University of Brasília (UnB), Faculty of Information Science, Brasília, Brazil
Amanda Damasceno de Souza, FUMEC University, Postgraduate Program in Information and Communication Technology and Knowledge Management
Eduardo Ribeiro Felipe, Federal University of Itajubá (UNIFEI), Institute of Technological Sciences, Itabira, Brazil

Contact: Fernanda Farinelli
email: fernanda.farinelli@unb.br

Published for:

14th International Conference on Biomedical Ontologies (ICBO 2023)

Conference website: https://www.icbo2023.ncor-brasil.org/index.html

Online publication:

CEUR Workshop Proceedings (CEUR-WS.org), http://ceur-ws.org

Preferred citation:

FARINELLI,F; SOUZA,A.D.; FELIPE, E.R. (eds.). ICBO 2023 – 14th International Conference on Biomedical Ontologies. *Proceedings of the 14th International Conference on Biomedical Ontologies, August 28th - September 1st, 2023.*Brasília, Brazil:CEUR-WS.org, 2023.

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

Acknowledgements

The organizers would like to thank everyone involved in making ICBO 2023 happen.

We would like to specially thank all authors and speakers for their contributions, and the Programme Committee members for their timely reviewing.

We thank the Faculty of Information Science (FCI) and the Central Library, both from the University of Brasilia, for their generous event support, organizational help, and also for providing their facilities.

Moreover, we would like to thank the International Society for Biocuration[1] for providing ISB Travel Fellowship support, and the Alma Sírio-Libanês[2] for sponsoring and organizing the panel.

We also thank the Brazilian Chapter of the National Center for Ontology Research (NCOR-BR[3]) for providing our website infrastructure. Last but not least, we acknowledge the generous hosting of the virtual event by Anthony Huffman and the University of Michigan Medical School.

Conference Host:

Administrative Support:

Organization Support:

[1] https://www.biocuration.org/
[2] https://alma.hsl.org.br/
[3] https://ncor-brasil.org/

ICBO 2023 Organization Committee

General Chair of the ICBO-Ontobras 2023 Joint Conference

Prof. Dr. Fernanda Farinelli (University of Brasília, Brazil)

Local Chair

Prof. Dr. Fernanda Farinelli (University of Brasília, Brazil)

Prof. Dr. Dalton Lopes Martins (University of Brasília, Brazil)

Program Committee Chairs

Prof. Dr. Jeanne Louize Emygdio (Pontifical Catholic University of Minas Gerais, Brazil)

Prof. Dr. Christiano Pessanha (Federal Fluminense University, Brazil)

Prof. Dr. Luis M. Machado (University of Coimbra, Portugal)

Workshops, Tutorials and Software Demo Chairs

Dr. Asiyah Yu Lin (Axle Informatics LLC, USA)

Dr. Andrey Soares (University of Colorado Anschutz Medical Campus, USA)

Journal of Biomedical Semantics ICBO Thematic Series Chairs

Prof. Dr. Stefan Schulz (Medizinische Universität Graz, Austria)

Prof. Dr. Amanda Damasceno de Souza (FUMEC University, Brazil)

Proceedings Chairs

Prof. Dr. Amanda Damasceno de Souza (FUMEC University, Brazil)

Prof. Dr. Eduardo Ribeiro Felipe (Federal University of Itajubá, Brazil)

Prof. Dr. Fernanda Farinelli (University of Brasília, Brazil)

Publicity Chair

Prof. Dr. Amanda Damasceno de Souza (FUMEC University, Brazil)

Website Chairs

Prof. Dr. Eduardo Ribeiro Felipe (Federal University of Itajubá, Brazil)

Dr. William Duncan (University of Florida, USA)

ICBO 2023 Program Committee

- Adrien Barton - The French National Centre for Scientific Research
- Alan Ruttenberg - The State University of New York at Buffalo
- Alexander D. Diehl - The State University of New York at Buffalo
- Amanda Damasceno de Souza - FUMEC University
- Andrey Soares - University of Colorado
- Asiyah Yu Lin - National Institutes of Health (NIH) & Axle Informatics LLC
- Barry Smith - The State University of New York at Buffalo
- Bjoern Peters - La Jolla Institute for Allergy and Immunology
- Carlos H. Marcondes - Federal Fluminense University
- Chris Stoeckert - University of Pennsylvania
- Christiano P. Pessanha - Federal Fluminense University
- Claudenir Fonseca - Free University of Bolzano
- Darren Natale - The Georgetown Faculty
- Eduardo Felipe - Federal University of Itajubá
- Fernanda Farinelli - University of Brasília
- Fernando Cruz - University of Brasília
- Flavio S Correa da Silva - University of São Paulo
- Fred Freitas - Federal University of Pernambuco
- Fumiaki Toyoshima - IRIT
- Gabriela Henning - INTEC (CONICET-UNL)
- Guylerme Figueiredo - Petrobrás S.A.
- Jeanne Louize Emygdio - Pontifical Catholic University of Minas Gerais
- Jeff Otte - The State University of New York at Buffalo
- Jie Zheng - University of Pennsylvania
- João Lima - Federal Senate of Brazil
- João Paulo A. Almeida - Federal University of Espírito Santo
- John Beverley - The State University of New York at Buffalo
- Jose M Parente de Oliveira - Aeronautics Technological Institute
- Lais Salvador - Federal University of Bahia
- Leo Obrst - MITRE
- Luan Garcia - Federal University of Rio Grande do Sul
- Luis M. Machado - University of Coimbra

- Luiz Olavo Bonino da Silva Santos - University of Twente
- Mara Abel - Federal University of Rio Grande do Sul
- Marcio Bezerra da Silva - University of Brasília
- Maria das Graças da Silva Teixeira - Federal University of Espírito Santo
- Mauricio B. Almeida - Federal University of Minas Gerais
- Phillip Lord - Newcastle University
- Pierre Grenon - National Center for Ontological Research
- Rafael Peñaloza - University of Milano-Bicocca
- Randi Vita - La Jolla Institute for Allergy & Immunology
- Renata Guizzardi - University of Twente
- Sana Debbech - IRT Railenium
- Sarah Alghamdi - King Abdullah University of Science and Technology
- Silvio Gonnet - Universidad Tecnológica Nacional/Facultad Regional Santa Fe
- William Duncan - University of Florida
- Yongqun “Oliver” He - University of Michigan

Pre-conference Workshops and Tutorials

Pre-conference workshop proposals accepted:

1. Ontologies for Infectious and Immune-mediated Disease Data Science, co-organized by Asiyah Yu Lin, Alexander Diehl, John Beverley, and Lindsay Cowell.
   Monday, August 28, 2023. Workshop page[4].

2. The 7th International Cells in Experimental Life Science Workshop, CELLS 2023, co-organized by Alexander D. Diehl and Yongqun "Oliver" He.
   Tuesday, August 29, 2023. Workshop page[5].

3. 12th Vaccine and Drug Ontology Studies (VDOS) 2023 Workshop, co-organized by Junguk Hur, Cui Tao, and Yongqun "Oliver" He.
   Tuesday, August 29, 2023. Workshop page[6].

4. FOHTI-23 (FAIR Ontology Harmonization and Trust Data Interoperability), co-organized by Anna Maria Masci and Asiyah Yu Lin.
   Tuesday, August 29, 2023. Workshop page[7].

Pre-conference tutorial proposals accepted:

1. Natural Language Processing Tutorial for Biomedical Text Mining, co-organized by Senay Kafkas, Sumyyah Toonsi, Sakhaa Alsaedi.
   Monday, August 29, 2023. Workshop page[8].

2. OBO Foundry Tutorial: Introduction to workflows and best practices for ontology development, co-organized by Nicole Vasilevsky, Nico Matentzoglu and Tiago Lubiana.
   Monday, August 29, 2023. Workshop page[9].

[4] https://delaneycdmcnulty.wixsite.com/oiids-workshop
[5] https://sites.google.com/view/cells2023/home
[6] https://vdos-workshop.github.io/vdos2023/
[7] https://fohti.github.io/FOHTI-2023/
[8] https://github.com/stoonsi/ICBO-NLP-for-Biomedical-Text-Mining-tutorial/tree/main
[9] https://oboacademy.github.io/obook/courses/icbo2023/

3. BFO as a top-level ontology for information systems modeling, co-organized by Maurício B. Almeida and Jeanne L. Emygdio.

   Tuesday, August 29, 2023. Workshop page[10].

4. Machine Learning with Ontologies, co-organized by Maxat Kulmanov, Robert Hoehndorf, Sarah Alghamdi, Azza Althagafi, Sumyyah Toonsi, Fernando Zhapa-Camacho.

   Tuesday, August 29, 2023. Workshop page[11].

5. Ontology-Driven Conceptual Modeling with UFO, gUFO, and OntoUML, co-organized by Giancarlo Guizzardi.

   Monday, August 29, 2023. Workshop page[12].

[10] https://mba.eci.ufmg.br/tutorialicbo2023/

[11] https://github.com/bio-ontology-research-group/mowl-tutorial

[12] https://www.inf.unibz.it/krdb/KRDB files/KRDB-SOS-2020-slides/KRDB-SOS-2020-Guizzardi.pdf

Software Demo

Title: Ontology Development in FHIR Resources with the Fast Evidence Interoperability Resources (FEvIR) Platform

Presenter: Caue Monaco

Tool link: https://fevir.net

Alma Sírio-Libanês Panel

Theme: Ontologies Application in the Brazilian Health Area

Coordinator: MD. Beatriz de Faria Leão

List of Panels:

- Panelist: MD. Fabio Cerqueira Lario (Sírio-Libanês Hospital, São Paulo)
- Panel title: The role of ontologies in electronic medical record systems

- Panelist: MD. Jussara Rötzsch Macedo
- Panel title: The contribution of SNOMED-CT to the Brazilian National Health Data Network (RNDS)

- Panelist: Dr. Joice Machado (SOFYA - Plug and Play Deep Medical AIs)
- Panel title: Use of SNOMED-CT sub-ontologies in the automatic identification of allergies and adverse reactions in clinical notes

- Panelist: Robson Mattos (CGIIS/SEIDIGI Digital/Ministry of Health of Brazil)
- Panel title: OBM - The Brazilian Ontology of Medicines

Preface

With great honor, we present the ***Proceedings of the 14th International Conference on Biomedical Ontology (ICBO 2023)***, convened at the Faculty of Information Science, University of Brasilia, Brasília, DF, Brazil. This year, the Conference Hosts was both the Undergraduate course in Library Science and the Postgraduate Program in Information Science of the Faculty of Information Science at the University of Brasília. The ICBO represents a preeminent annual conference series that convenes researchers, scholars, and professionals engaged in the development and application of ontologies across biomedical domains and related fields.

This edition of the conference marked a historic moment, being the first time this significant event occurred in Latin America, within Brazil's borders. This edition took place in conjunction with the 16th Ontology Research Seminar in Brazil (ONTOBRAS 2023) as a satellite conference. This joint conference served as a premier forum, facilitating the intellectual exchange of groundbreaking ideas and profound discoveries among the ontology community.

Both ICBO and ONTOBRAS series are consolidated and independent events, which have already established themselves as reference events in the field of applied ontology research, the first at an international level dedicated to biomedical ontologies and the second as nationally (Brazilian) known internationally and dedicated to ontologies in general. These distinct yet interlinked gatherings persist as vanguards, setting the standard for ontology research, spanning theoretical constructs to practical applications across diverse domains of knowledge. This volume of proceedings focuses only on the ICBO program, as ONTOBRAS 2023 conceives its own proceedings in the IAOA Series.

Biomedical ontology plays a central role in the development of health information systems, clinical data integration, and translational research facilitation, serving as a crucial component of artificial intelligence and machine learning solutions. Ontologies are not merely intellectual tools; they are foundational elements enabling effective communication among researchers, clinicians, and information systems by providing unambiguous and high-quality artifacts of knowledge representation. They empower the discovery of patterns, the analysis of complex data, and the creation of innovative applications that directly benefit human health, while also driving advancements in artificial intelligence and machine learning within the biomedical context.

The overarching theme that directed the ICBO 2023 program was "***The Role of Ontologies in Artificial Intelligence and Machine Learning***." In addition to the scientific paper presentation sessions, the conference encompassed four workshops, five tutorials, a panel dedicated to the dissemination of practical experiences from Brazil, a session for presenting posters, and a software demonstration, all meticulously integrated into its program. Furthermore, the conference featured four keynote lectures of eminent significance. These four outstanding keynote presentations featured at the conference were as follows:

- Dr. Mathias Brochhausen: "Improving Ontologies to Foster AI in Biomedicine"
- Dr. Maurício B. Almeida: "A BFO-Based Authority Framework for Healthcare Corporations"
- Dr. Barry Smith: "Use of AI in Medical Research and the Role of Ontology"
- Dr. Renata Guizzardi: "Ethical Requirements for AI Systems"

We previously presented detailed information about the workshops, tutorials, software demos and discussion panels.

ICBO 2023 successfully took place in a hybrid format, accommodating over ***40 in-person attendees and 63 virtual participants***. Additionally, during our joint session - workshops, tutorials, keynotes, poster/software demo session, and panel, we welcomed **52 in-person attendees from Ontobras**. This conference provided a diverse array of research endeavors and practical applications. From theoretical explorations that push the boundaries of knowledge to pragmatic implementations enhancing healthcare quality, ICBO participants exhibited a profound dedication to academic and scientific excellence.

For the main conference, 24 papers and abstracts underwent rigorous peer review by three members selected by the Program Committee Chairs. We accepted 20 of these submissions for inclusion in this volume, including 14 full papers, 2 short papers, and 4 poster abstracts. We also included in this publication abstracts of the tutorials and the software demonstration, as well as contributions from the Workshop on Ontologies for Infectious and Immune-Mediated Disease Data Science (OIIDDS 2023) and the FAIR Ontology Harmonization and Trust Data Interoperability Workshop (FOHTI 2023). We extend our sincere gratitude to the program committee members for their discerning evaluations, which significantly contributed to the quality of this compilation.

As readers delve into the proceedings of this event, we encourage profound engagement with the presented themes, urging critical inquiry into preconceptions and established paradigms. Each paper and abstract featured herein adds to the ongoing advancement of biomedical ontology, fostering innovation and collaboration in our collective pursuit of a healthier, more enlightened future.

We would like to express our gratitude to the authors for their hard and quality work, the reviewers for their dedicated time and expertise, and the organizers for their diligence in putting on this conference.

We also would like to express our heartfelt gratitude to the distinguished keynote speakers for their exceptional and inspiring presentations, which greatly enriched the intellectual discourse at ICBO 2023. Finally, but not less importantly, we extend our heartfelt gratitude to the organizers and presenters of the tutorials and workshops for their exceptional efforts and invaluable contributions, which greatly enriched the conference experience for all attendees.

In closing, we hope that this conference’s proceedings will serve as inspiration for future work, fuel innovative pursuits, and strengthen connections, propelling our collective efforts to new heights and ensuring vibrant discussions at upcoming conferences within our esteemed ontology community.

October, 2023

Fernanda Farinelli
Amanda Damasceno de Souza
Eduardo Ribeiro Felipe

Table of Contents

Session 1: Main Conference

I - Full papers:

II - Short papers:

III - Poster abstracts:

Session 2: Workshop, Tutorials and Software Demo

I - Workshop on Ontologies for Infectious and Immune-Mediated Disease Data Science (OIIDDS 2023)

II - Workshop on FAIR Ontology Harmonization and trust Data Interoperability (FOHTI 2023)

III - Tutorial Abstracts

IV - Software Demo Extended Abstract

Ontology-based representation and analysis of conditional vaccine immune responses using Omics data

Anthony Huffman[1], Edison Ong[12], Tim Brunson[3], Nasim Sanati[3], Jie Zheng[4], Anna Maria Masci[5], Guanming Wu[3], Yongqun He[6-8]

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States.
[2] GlaxoSmithKline, Rixensart 1330, Belgium.
[3] Oregon Health & Science University, Portland, Oregon, OR, United States.
[4] Department of Genetics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, United States
[5] Office of Data Science National Institute of Environmental Health Science, Durham, NC, United States.
[6] Unit of Laboratory Animal Medicine, University of Michigan, Ann Arbor, MI, United States.
[7] Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI, United States.
[8] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States.

Abstract

ImmPort, the world's largest repository of immunology data, includes many vaccine immune response datasets. ImmPort maps the metadata of these studies to ontology and database schema. As of February 28, 2023, our ImmPort data analysis identified 6.258 immune exposures using 47 vaccines in 4,607 human subjects, and 324 cohort studies from the ImmPort. We hypothesized that an integrative ontological representation of the data from these studies would enhance our understanding and analysis of these ImmPort vaccine studies, and with ontological classification and tools such as VIGET, we could further study the effects of different conditions such as vaccine types and host biological sex on the vaccine response gene expression profiles. Our Vaccine Ontology (VO) analysis classified these 37 vaccines into bacterial, viral, and protozoan vaccine types with different vaccine properties. The ImmPort metadata types were modeled with the Vaccine Investigation Ontology (VIO). Our new ontology-based pipeline extracted vaccine response data from the ImmPort database, annotated them based on ontology, obtained corresponding gene expression data from the GEO, and performed consistent omics data analysis. Our use case found gene profiles shared and differed from live and killed inactivated influenza vaccines. Furthermore, our Omics data analysis using the VIGET tool found that female and male human subjects have differential host responses for influenza vaccines. For example, our study showed much stronger early female responses to influenza vaccination than males, and males was able to show active immune responses at a later stage. Interestingly, the female (but not male) human subject group also showed significantly enriched neutrophil degranulation at Day 3 after influenza vaccination; however, males (but not females) displayed significantly enriched neutrophil degranulation at Day 14 after influenza vaccination. These mechanisms have been used to find differences between the gene lists and pathways of host responses to different vaccines conditional to different factors including vaccine types and host biological sex. Moreover, this framework can be expanded to other vaccines and vaccine categories easily.

Keywords

Vaccine Ontology, ImmPort, GEO, Vaccine immune response, Gene expression profiles

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: huffmaar@umich.edu (A. 1); edong@umich.edu (A. 2); tjbrunson@gmail.com (A. 3), nasim@plenary.org(A.4), jiezhen@pennmedicine.upenn.edu(A.5), mascia2@nih.gov(A.6), wug@ohsu.edu(A.7) yongqunh@med.umich.edu (A.8)
ORCID: 0000-0002-5735-7540 (A. 1); 0000-0002-5159-414X (A. 2); 0000-0001-6681-0418 (A.4) 0000-0002-2999-0103 (A.5), 0000-0001-9189-9661 (A.6), 0000-0001-8196-1177 (A.7), 0000-0001-9189-9661 (A.8)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

As one of the greatest inventions in modern medicine, vaccination has been used to dramatically protect humans against infectious diseases and improve human health. However, infectious diseases are still a major cause of human mortality throughout the world, and effective vaccines that protect against many diseases still do not exist [1,2]. The future success of effective vaccine development relies on deep understanding of the protective immune mechanisms [3].

Vaccine induced host responses depend on various conditions and factors. For example, we have previously developed the VaximmutorDB, a web-based database system that has included over 1,700 experimentally identified vaccine immune effectors (abbreviated as "vaximmutors") induced by 154 vaccines for 46 pathogens [4]. The VaximmutorDB data have been manually annotated from peer-reviewed publications and reliable databases. Our VaximmutorDB data analysis showed that these vaccines induce many common immune factors, for example, Th1 immune factors IFN-gamma and IL-2 and Th2 immune factors IL-4 and IL-6. Responses induced by different vaccine types such as influenza and yellow fever vaccines may also differ [4]. However, more specific mechanisms underlying vaximmutors and vaccine immune responses are still largely unclear, even when we provided a successful use case for this yellow fever and influenzas vaccines for identifying these vaximmutors [4]. Numerous other studies also identified many other factors crucial for vaccine response induction [5-7]. For example, biological sex and age may change the immune responses significantly [7, 8].

ImmPort (the Immunology Database and Analysis Portal; https://ImmPort.niaid.nih.gov/) is the world's largest repository of public-domain de-identified clinical trial data related to immunology [9]. All data derived from clinical trials funded by the Division of Allergy, Immunology and Transplantation (DAIT) of the National Institute of Allergy and Infectious Diseases (NIAID) are required to be published on the ImmPort portal. In addition, the ImmPort portal includes data obtained from the work of the Human Immunology Project Consortium (HIPC, http://www.immuneprofiling.org/) as well as relevant data from several external sources such as the Gates Foundation. ImmPort includes complete clinical and mechanistic study data (e.g., mechanistic assays used, timing of visits, etc.), all of which are publicly available for download in a de-identified form.

Pathway/network based analyses are extremely important in understanding protective immune responses to infectious diseases. The immune systems of two vaccinees with different genetic backgrounds may have different gene expression profiles in response to the same vaccine; however, there may be shared underlying core immunological pathways between these two vaccinees. Various experimental conditions may also affect host immune responses to vaccines. Nevertheless, appropriate tools to support the analysis and visualization of vaccine-induced immune pathways under different given conditions still miss.

We hypothesize that different vaccine types induce differential but coherent host responses, and experimental conditions would significantly change the results of vaccine-induced host responses. More specifically, we wanted to see if by leveraging ontology, that we could integrate multiple vaccine studies to find common patterns for vaccines with the shared experiments. We have previously studied how different factors would affect the host responses to Yellow fever vaccines [10], influenza vaccines [11], and Brucella vaccines [12]. However, the previous studies were still limited with small datasets and deep investigations. In current study, we would investigate further how specific factors such as biological sex and vaccine types would change the vaccine-induced host responses.

Biomedical ontologies have emerged to be critical for the standardization, integration, and analysis of the large amounts of heterogeneous biological data. We previously demonstrated how ontologies, including the Vaccine Ontology (VO) [13], Vaccine Investigation Ontology (VIO) [14], and Ontology of Biological and Clinical Statistics (OBCS) [3, 15] could be used to support vaccine induced host response studies. Vaccine Ontology is a reference ontology focused on vaccines in general while Vaccine Investigation Ontology is an extension that is focused on metadata types in various vaccine investigation studies. Such approaches may be used to study the ImmPort data. Given the ImmPort database with its own data management system, it would be important to integrate our ontology-based approaches with the ImmPort data system for better studies.

VIGET is our newly developed web-based Vaccine response Gene Expression analysis Tool based on Reactome and ImmPort [16]. VIGET uses the VO to classify various vaccine types and experimental conditions. VIGET allows users to select vaccines with VO classification, choose ImmPort studies and confounding variables, and perform differential gene expression analysis of various vaccine responses. It is possible to apply VIGET to investigate the vaccine response gender differences by comparing patterns for gender-based differences across influenza vaccines. We want to leverage our ontology driven design to identify further patterns contained in multiple vaccine studies.

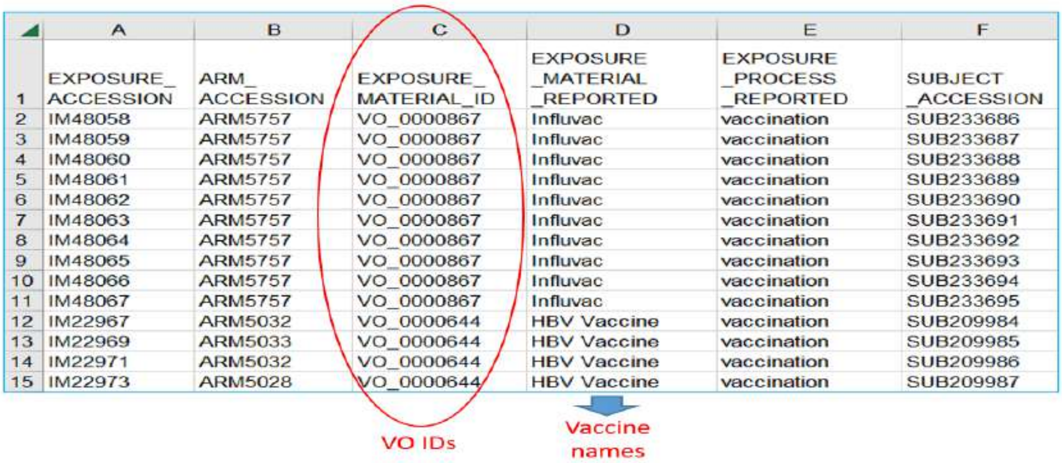
In this report, we demonstrated how ontology and VIGET can be applied to systematically standardize, classify, integrate, and analyze vaccine-related high-throughput ImmPort and GEO data, and identify vaccine-induced pathways under specific experimental conditions including vaccine types and biological sexes of the human subjects. Three specific use cases were also developed to further illustrate the effectiveness of our ontology-based representation and analysis of vaccine-induced host responses given different vaccine or host conditions.

2. Methods

2.1. ImmPort vaccine metadata extraction and storage

The immune exposure metadata description of various vaccine studies was downloaded from the public ImmPort website (https://www.immport.org/) to a local computer. Such metadata description is stored in different ways in ImmPort. ImmPort provides the annotations to the gene expression data that were deposited in GEO. The raw data can be downloaded from GEO. Our study downloaded the csv text file of such metadata, and converted it to Excel for easy exploration and processing.

The ImmPort database has the immune exposure table that stores the information on how subjects were immune exposed. Figure 1 shows how the vaccine immune exposure data in ImmPort was standardized using the Vaccine Ontology (VO). From the data in ImmPort, we added additional columns to create the metadata file via inter-rater discussion.



	A	B	C	D	E	F
1	EXPOSURE_ ACCESSION	ARM_ ACCESSION	EXPOSURE_ MATERIAL_ID	EXPOSURE _MATERIAL _REPORTED	EXPOSURE _PROCESS _REPORTED	SUBJECT _ACCESSION
2	IM48058	ARM5757	VO_0000867	Influvac	vaccination	SUB233686
3	IM48059	ARM5757	VO_0000867	Influvac	vaccination	SUB233687
4	IM48060	ARM5757	VO_0000867	Influvac	vaccination	SUB233688
5	IM48061	ARM5757	VO_0000867	Influvac	vaccination	SUB233689
6	IM48062	ARM5757	VO_0000867	Influvac	vaccination	SUB233690
7	IM48063	ARM5757	VO_0000867	Influvac	vaccination	SUB233691
8	IM48064	ARM5757	VO_0000867	Influvac	vaccination	SUB233692
9	IM48065	ARM5757	VO_0000867	Influvac	vaccination	SUB233693
10	IM48066	ARM5757	VO_0000867	Influvac	vaccination	SUB233694
11	IM48067	ARM5757	VO_0000867	Influvac	vaccination	SUB233695
12	IM22967	ARM5032	VO_0000644	HBV Vaccine	vaccination	SUB209984
13	IM22969	ARM5033	VO_0000644	HBV Vaccine	vaccination	SUB209985
14	IM22971	ARM5032	VO_0000644	HBV Vaccine	vaccination	SUB209986
15	IM22973	ARM5028	VO_0000644	HBV Vaccine	vaccination	SUB209987

Figure 1: ImmPort immune exposure metadata illustration. The immune_exposure.txt file was downloaded on March 28, 2023, from ImmPort website, opened in Excel format, and the vaccine related information extracted. Note that it is only a portion of the contents in the file, and some columns of the file are not shown here to save space. We added The Vaccine Ontology (VO) is used here for vaccine information standardization.

2.2. Vaccine response data analysis using ontology-annotated ImmPort and GEO

We used Ontofox [17] to generate a hierarchy of vaccine terms from VO that mapped to vaccine terms studied in different experiments from ImmPort. We used the Vaccine Investigation Ontology (VIO) to map and model metadata types for the vaccine experiments. Protégé-OWL editor was used for ontology display and editing. SPARQL and DL-Query were used for ontology knowledge query.

2.3. Vaccine response data analysis using ontology-annotated ImmPort and GEO

ImmPort provided vaccine information and the metadata for different vaccine response studies. The vaccine types and metadata were annotated using VO and VIO. Based on the metadata of different studies, we extracted the normalized microarray gene expression data from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). Software programs and APIs were generated to automatically query the final integrative ontology and ImmPort data. As a use-case, influenza vaccine studies were selected through ontology-based queries along with ontology-defined vaccine and condition (e.g., health and age) classification. For GEO data analysis, log2 transformation was applied if the expression values of a given GEO dataset were not in log space. For each comparison pair, expression data collected from individuals vaccinated on Day 0 was considered as the control group, while data collected on Day 3 to Day 7 after vaccination were considered as the vaccine exposure group. In addition, each pair was defined to have the same ontology-defined vaccine and condition within the same cohort in the same ImmPort study.

2.4. Omics data analysis using VIGET

The web-based Reactome pathway analysis tool (https://reactome.org/) was used to support our gene expression profile analyses. Adjusted FDR p-value cutoff of 0.05 was used for statistical data analysis [18].

3. Results

3.1. Vaccine response data extraction from ImmPort

The ImmPort website maintains a publicly available data model schema (https://immport.org/shared/dataModel) and a corresponding relational database. ImmPort does not store raw Omics gene expression data. However, ImmPort provides the annotations to the gene expression data that were deposited in GEO. The raw data can be downloaded from GEO.

An important table in ImmPort database is the immune exposure table that stores the information on how subjects were immune exposed. Figure 1 shows how the vaccine immune exposure data in ImmPort was standardized using the Vaccine Ontology (VO). Basically, the metadata file has many columns including Exposure Accession, ARM Accession, Exposure Material ID (which uses VO IDs for vaccines), Exposure Merial Reported (which is the label of the material), Exposure Process Reported (which is “vaccination” for vaccine immune exposure studies), and Subject Accession. With the VO IDs used in ImmPort metadata file, it became easy and efficient for us to standardize and process the vaccine data and categorization.

Based on our analysis, as of February 28, 2023, the ImmPort database included 36,140 immune records. Our analysis of the database identified 6,258 vaccine-related immune exposure records that used 37 vaccine terms with unique VO IDs in 4,607 human subjects, 324 ARMs (a single group or cohort for a specific study purpose).

3.2. VO-based classification of the ImmPort-reported vaccine studies

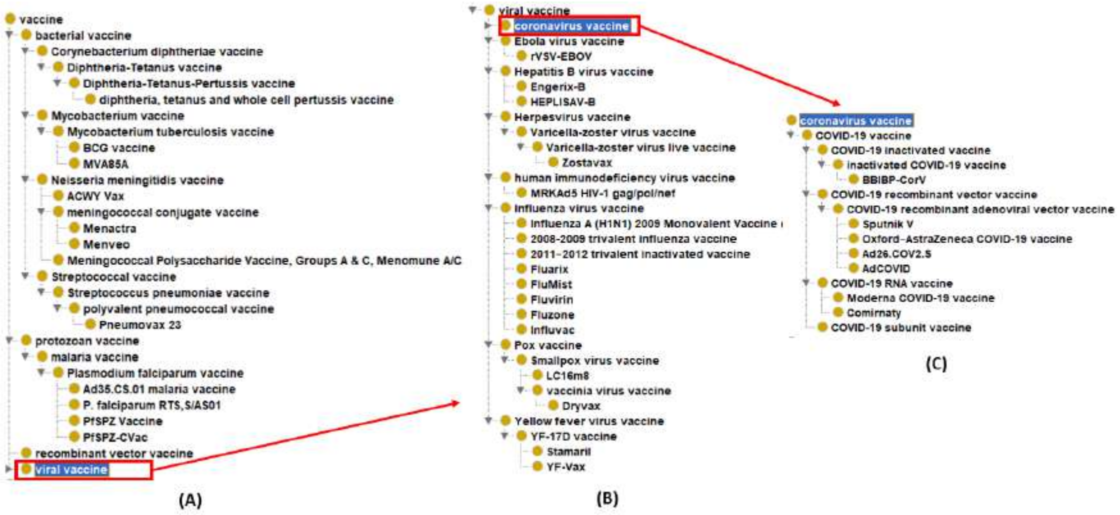

vaccine
bacterial vaccine
Corynebacterium diphtheriae vaccine
Diphtheria-Tetanus vaccine
Diphtheria-Tetanus-Pertussis vaccine
diphtheria, tetanus and whole cell pertussis vaccine
Mycobacterium vaccine
Mycobacterium tuberculosis vaccine
BCG vaccine
MVA85A
Neisseria meningitidis vaccine
ACWY Vax
meningococcal conjugate vaccine
Menactra
Menveo
Meningococcal Polysaccharide Vaccine, Groups A & C, Menomune A/C
Streptococcal vaccine
Streptococcus pneumoniae vaccine
polyvalent pneumococcal vaccine
Pneumovax 23
protozoan vaccine
malaria vaccine
Plasmodium falciparum vaccine
Ad35.CS.01 malaria vaccine
P. falciparum RTS,S/AS01
PfSPZ Vaccine
PfSPZ-CVac
recombinant vector vaccine
viral vaccine
(A)
viral vaccine
coronavirus vaccine
Ebola virus vaccine
rVSV-EBOV
Hepatitis B virus vaccine
Engerix-B
HEPLISAV-B
Herpesvirus vaccine
Varicella-zoster virus vaccine
Varicella-zoster virus live vaccine
Zostavax
human immunodeficiency virus vaccine
MRKAd5 HIV-1 gag/pol/nef
influenza virus vaccine
Influenza A (H1N1) 2009 Monovalent Vaccine
2008-2009 trivalent influenza vaccine
2011-2012 trivalent inactivated vaccine
Fluarix
FluMist
Fluvirin
Fluzone
Influvac
Pox vaccine
Smallpox virus vaccine
LC16m8
vaccinia virus vaccine
Dryvax
Yellow fever virus vaccine
YF-17D vaccine
Stamaril
YF-Vax
(B)
coronavirus vaccine
COVID-19 vaccine
COVID-19 inactivated vaccine
inactivated COVID-19 vaccine
BBIBP-CorV
COVID-19 recombinant vector vaccine
COVID-19 recombinant adenoviral vector vaccine
Sputnik V
Oxford-AstraZeneca COVID-19 vaccine
Ad26.COV2.S
AdCOVID
COVID-19 RNA vaccine
Moderna COVID-19 vaccine
Comirnaty
COVID-19 subunit vaccine
(C)

Figure 2: Ontological representation of 37 vaccines in ImmPort and their relations using VO. Vaccines in VO are asserted by the pathogen they vaccinate against. Column A shows that ImmPort contains a small selection of bacterial vaccines (7) and parasite vaccines (4). 3 of the 37 vaccines were mapped to 2008-2009 influenza vaccine. The red boxes represent parent terms shown later in the hierarchy.

3.3. VIO-based of the vaccine investigation conditions

Figure 3 illustrates the basic flow to extract GEO expression data accession in the ImmPort database schema by following six primary accessions. The flow chart represents the annotations under each GEO GSM ID. Each study has one or more ARM (or cohort). Each ARM has one or more subjects. Each subject has one or more biosample. Each biosample has one or more experimental sample (i.e., expsample). Meanwhile, each subject also has an immune exposure such as vaccination, and each vaccination exposure points to a VO ID. Finally, each experimental sample corresponds to a specific GSM ID. Each table also has more detailed annotations in the ImmPort database. Eventually, the GSM ID can be used to get the data from the GEO database (Figure 3).

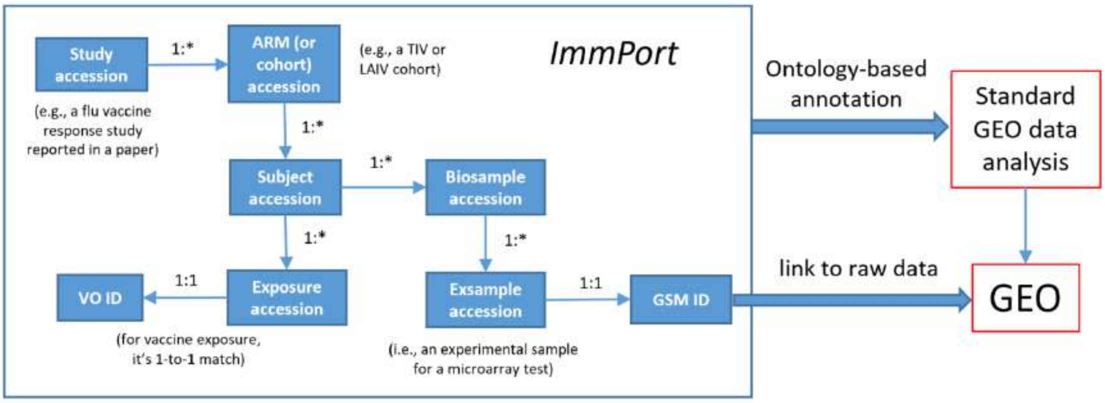

ImmPort
Study accession
(e.g., a flu vaccine response study reported in a paper)
1:*
ARM (or cohort) accession
(e.g., a TIV or LAIV cohort)
1:*
Subject accession
1:*
Biosample accession
1:*
1:*
VO ID
1:1
Exposure accession
(for vaccine exposure, it's 1-to-1 match)
Exsample accession
(i.e., an experimental sample for a microarray test)
1:1
GSM ID
Ontology-based annotation
Standard GEO data analysis
link to raw data
GEO

Figure 3: Vaccine response Omics data analysis using ontology-annotated ImmPort and GEO. Six IDs in the boxes represent the primary keys of corresponding tables of the ImmPort database. These six

primary accessions are IDs for the study (study accession), cohort of a study (ARM accession), subject of a cohort (subject accession), vaccine used on a subject (exposure accesion), biological test used on a subject (biosample accession), the biological assay utilized on the subject (biosample accession), and the experimental sample that was part of the biological assay (exsample accesion). The schema illustrated here explains the components in Figure 1 and the relations among them, and links the ImmPort data to GEO data.

3.4. Use case 1: Ontology-based query of vaccines and vaccine investigation data

Based on the logically defined hierarchies and semantic relations in ontology, we can use our ontology system to perform computer-assisted queries and analysis.

For example, based on the VO ontology hierarchy as shown in Figure 2, we can easily identify which vaccines are bacterial vaccines, and which are viral vaccines. We can also detect which are influenza vaccines or yellow fever vaccines. VO classification clearly lays out the relations of these vaccines. Since the VO is computer-interpretable, the results can be automatically parsed by a semantic query such as SPARQL or DL-query for various follow-up analyses or to aid retrieval of experimental studies as part of use case 2.

Not shown in Figure 2, VO includes axioms that illustrate different properties of vaccines. For example, the FluMist vaccine is defined to have live attenuated feature as seen in the following axiom:

'has quality' some 'vaccine organism live attenuated'

In contrast, the Fluarix vaccine is defined to be an inactivated whole organism vaccine based on the following axiom:

'has quality' some 'vaccine organism inactivated'

Such logical axioms defined in VO provide us a logical way to automatically identify different features. For example, we can use SPARQL or DL-query to easily find which vaccines are live attenuated vaccines and which are inactivated vaccines. This was done to retrieve experimental studies for further analysis of the vaccines within ImmPort.

Note that the ImmPort database does not provide the semantic information described above. We extracted the VO IDs annotated in ImmPort and generated a VO subset that contains these VO IDs and their corresponding vaccines and related vaccine attributes. This subset of VO was then applied to support our semantic queries and analysis.

3.5. Use case 2: Detecting the effect of biological sex on gene expression profiles stimulated by influenza vaccines

After we identify which vaccines or vaccine groups to evaluate, we can come back to query the ImmPort database to obtain detailed vaccine study information (Figure 2). For example, using the ontology strategy, we identified many vaccines grouped as inactivated influenza vaccines as shown in Figure 2. Using ImmPort database, we identified three studies involving Fluzone and Fluarix, two inactivated influenza vaccines which will be linked to GEO data as part of use case 2. It is known that males and females may have different responses to vaccination. To see if using VIGET can recapitulate this vaccine response gender difference, we compared patterns for gender-based differences across influenza vaccines. We expanded our analysis to include non-immune related pathways related to cell transcription or cell signaling. For influenza vaccines, we utilized all 16 influenza studies covering one attenuated influenza vaccine, FluMist (VO_000044), and five inactivated influenza vaccine, Fluarix (VO_000044), Fluzone (VO_000047), Fluvirin (VO_000046), the 2008-2009 trivalent inactivated vaccine (VO_0004808) and the 2011-2012 trivalent inactivated vaccine (VO_0004810). These include the same vaccines as part of use case 2. Our VIGET study included 210 male and 260 female subject samples collected from either whole blood or peripheral blood mononuclear cells (PBMCs) [16]. These subjects ranged from 0 years old to 90 years old and included the same ethnicities as the previous use case. Finally, due to reduced sample sizes after 14 days, our analysis is focused on days 3, 7, and 14 in comparison to day 0. All results were adjusted by age, race, and vaccine type. Using the VIGET tool, we analyzed the effect on pathways using differences in log10-fold expression greater than 0.1, 0.2, 0.5, 1.0. These pathways can be found as part of Supplemental Table 1.

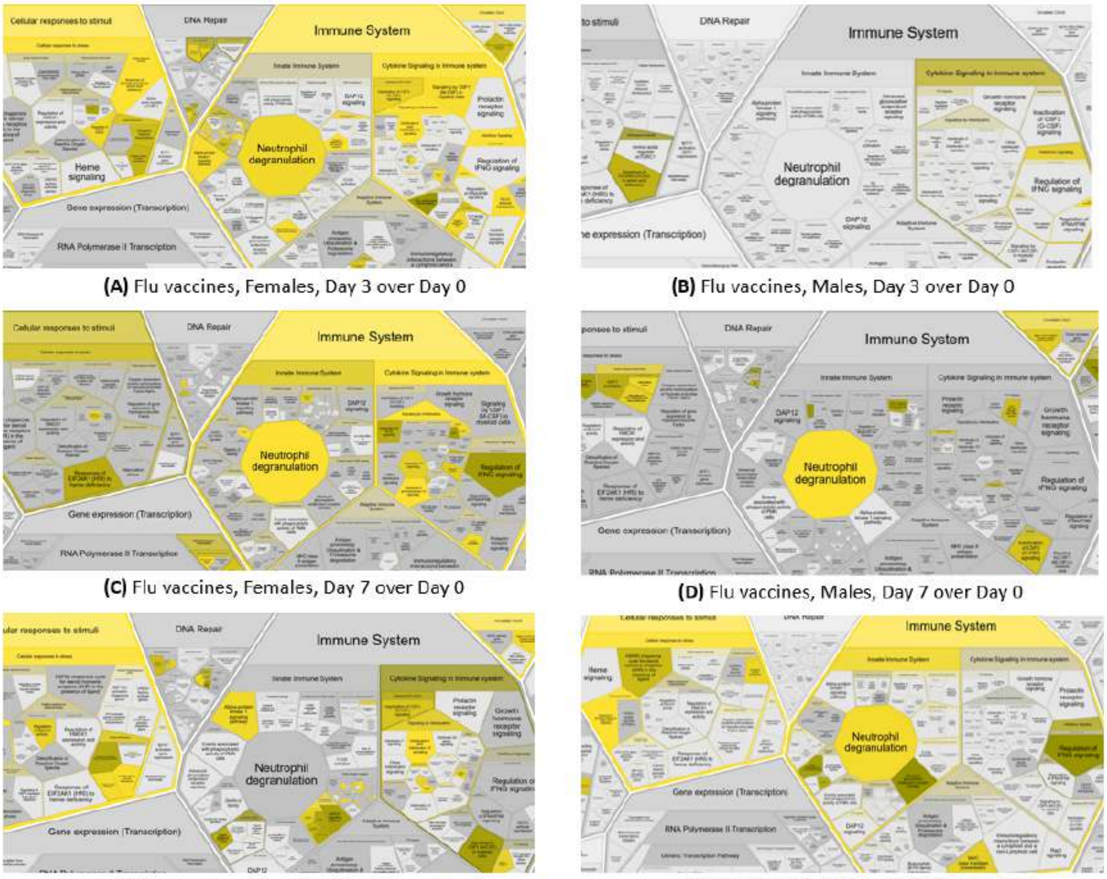
Our study showed that female vaccine responses tended to exhibit greater fold change than males earlier to the time response. Using the default 0.2 pathway reveals that males exhibited less significant pathways in females at all 3 time points (Table 1). Day 7 had the highest amount of significant genes and pathways for both males and females. For males, significant pathways only emerged at log-fold changes of 0.5 or greater; with the 1 log-fold expression showing neutrophil degranulation (FDR = 6.48 e-7) and innate immune system (FDR = 1.02e-2) as the only significant pathways. Females, in contrast, had 10 pathways at the same time point (Table 1), including multiple immune response pathways, including neutrophil degranulation (FDR = 5.06 e-11), IL-4 and IL-13 signaling (FDR = 2.31e-5), IL-10 signaling (FDR = 4.01e-2), and the CLEC7A/inflammasome pathway (FDR = 4.77e-2). Additionally, females at day 7 also exhibited cellular response to stress (FDR = 3.19e-4), with all genes in this pathway being up-expressed (Supplemental Table 1, Flu-F-07-1.0-Pos.) Otherwise, there is a consistent pattern of males having fewer significant pathways, and immune related pathways than females for influenza vaccines.

Table 1. Summary of gender differences in number significant genes and pathways in influenza vaccine response. Differential gene response for influenza vaccines across males and females. The initial timepoint for comparison was Day 0 for all entries. All genes that had the magnitude of their log fold change greater than the threshold were counted. For # of significant pathways, the numbers indicate the number of pathways with FDR < 0.05. The full list of pathways can be found as part of the Supplemental Table 1. Day 14 Males had no pathways related to immune response. All log-fold changes are base 10. Results were corrected for age, race, batch, platform.

Time Comparison	Log-Fold Change Threshold	# of Significant Flu M Genes	# of Significant Flu M Pathways	# of Significant Flu F Genes	# of Significant Flu F Pathways
Day 3	0.1	176	17	969	73
	0.2	3	35	56	126
	0.3	0	0	9	126
	0.5	0	0	0	0
Day 7	0.1	1388	0	1427	0
	0.2	1329	0	1052	3
	0.3	1133	0	1048	0
	0.5	1119	0	690	8
	1.0	626	2	515	10
Day 14	0.1	446	39	969	73
	0.2	44	36	98	70

	0.3	4	68	11	34
	0.5	0	0	0	0

Figure 4 shows the Reactome pathway enrichment analysis results at three days, seven days, and fourteen days post-vaccination for males and females. Overall, females showed a significantly earlier immune responses at Day 3 post influenza vaccination than males, and then later males caught up with active immune responses. As shown in this figure, females exhibit an earlier vaccine response as shown in pathways that are linked to signaling of interleukins 4, 13 and 10. An interesting finding is the difference in neutrophil degranulation pathway expression in influenza-vaccinated female and male groups. At Day 3 neutrophil degranulation was significantly enriched in the influenza-vaccinated female group but not in the male group. At Day 7, both female and male groups showed significant enrichment of neutrophil degranulation. However, in Day 14, only the male (but not female) group showed significant enrichment of neutrophil degranulation (Figure 4).



(A) Flu vaccines, Females, Day 3 over Day 0

(B) Flu vaccines, Males, Day 3 over Day 0

(C) Flu vaccines, Females, Day 7 over Day 0

(D) Flu vaccines, Males, Day 7 over Day 0

(E) Flu vaccines, Females, Day 14 over Day 0

(F) Flu vaccines, Males, Day 14 over Day 0

Figure 4: Comparison of sex-based differences in immune response to influenza vaccines. All Reactome subfigures were compared to Day 0 of vaccine administration. The conditions for subfigures were labeled with text in the figure. Subfigures were generated by Reactome's Reacfoam tool (Release 82, September 2022). The detailed information is provided in Supplemental Table 1.

We also investigated for sex differences that are common to live attenuated and inactivated influenza vaccines. Due to a small data set for female inactivated influenza vaccine users, analysis for Day 0 to Day 3 females could not be done using VIGET (Supplemental Table 2). However, inactivated influenza vaccines revealed patterns of increased ribosomal translation as being enriched at Day 3 with fewer immune pathways for males and females. Day 7 for the influenza vaccines failed to find pathways that were found significant in males. Day 14 revealed 44 enriched immune related pathways. Males had uniquely enriched pathways linked to the general immune system and interferon alpha/beta signaling (Supplemental Table 2). Females, in contrast, had enriched pathways linked to innate immunity enriched but not the general immune system. Intriguingly, day 7 and day 14 both reveal that females have enriched pathways related to programmed cellular death and other cellular functions. Further investigation of yellow fever vaccines revealed a similar pattern of additional enrichment of cell death and cell signaling. Taken together, this shows that females have a common, unique response to the vaccines we have tested.

4. Discussion

4.1. Application of Ontologies to ImmPort

The contributions of this study are multiple. First, we demonstrate our application of ontologies and semantic relations to annotate the ImmPort data and link to GEO database. Second, we studied the effect of vaccine type as a vaccine factor on the vaccine immune response profiles. Third, we detected the effect of biological sex as an important host factor that affects the vaccine immune responses.

Our study shows that ontology can be applied with the relational ImmPort database to better support vaccine immune response data analysis. There were initial difficulties with finding an appropriate mapping between ImmPort and GEO due to the multiple lookup tables from each study to the list of GEO IDs. While this was aided by ImmPort documentation, it still took significant effort trawling through the database to find the matching IDs. There were also issues resolving if time matching definitions for day 0 meant that data was taken before or after vaccination. However, once done, we can now use this framework to expand to other vaccines or vaccine categories in ImmPort. As such, this approach satisfies and aids in data FAIRness (Findable, Accessible, Interoperable, and Reusable). While adaption of new axioms or ontology terms do require manual annotation, these can be easily done following eXtensible Ontology Design (XOD) principles [19] and reusing a suite of Ontozoo tools designed for this task [20, 21].Two database systems can be applied simultaneously. The ontology triple store can be used to store ontology knowledge and metadata, and the relational database can be used to store instance data. We can then use different query languages to link them together seamlessly. This can be adapted to other databases, albeit after looking to find the best mapping for data. The use of ontology can then be used to aid further intersections of different vaccination procedures (vaccine types, vaccine timing, or different vaccination routes) and patient phenotypes (sex or age-based differences between vaccine response). Ontology standardization has already been suggested to ImmPort to standardize different cell types [22]. However, this was focused on mapping terms to biological entity terms from Gene Ontology, Cell Ontology, Protein Ontology and the Ontology of Biomedical Investigation. The use of ontology standardization for vaccines on these datasets is novel outside of our lab.

It is to the best of our knowledge, that this is the first time that both databases were linked together to gain information on the gene expression profile differences of TIV and LAIV. As GEO and ImmPort have different focuses, this linkage will help facilitate greater understanding of the data. According to a PubMed search using “GEO ImmPort database” (8/16/2021), GEO and ImmPort, have been used as data sources to identifying genes related to cancer development and survival [23, 24].

4.2. The Effect of Biological Sex on Vaccine Immune Response for ImmPort vaccines

Our study illustrates the significant effect of biological sex in the vaccine immune response generation. The analysis of influenza vaccines showed females having an earlier and much stronger activation of immune related pathways than males during the first week of vaccination. Moreover, female immune responses at Day 7 uniquely had all genes part of cellular response to stress be upregulated. It has been reported that females have a stronger early immune response than males to vaccines [25], with females more likely to experience adverse events caused from an autoimmune response, which may explain why these traits were only found when looking only at females. As this pattern was found across live attenuated and dead influenza vaccines, this may be the result of vaccines not being as optimized for human females and represents an avenue for further vaccine improvement and mechanism research. It is interesting to find the significant differences in neutrophil degranulation pathway expression in influenza-vaccinated female and male groups (Figure 6). The influenza-vaccinated females induced fast neutrophil degranulation at the early stage (Day 3) than males. At Day 7, both female and male groups showed similar enrichment of neutrophil degranulation. However, only males (but not females) had significant enrichment of neutrophil degranulation at Day 14 post vaccination (Figure 6). Neutrophil degranulation is the regulated exocytosis of secretory granules containing cellular mediators such as proteases and inflammatory mediators. Many recent studies infer various roles of neutrophil degranulation in inflammatory disorder (e.g. septic shock and asthma) [26], bacterial virulence strategy [27], and COVID-19 infection [28]. However, the role of neutrophil degranulation in vaccine immune response induction is still unclear and worth further investigation.

Still, there are other studies that validate our results. A prior article that also identified increases of ZAP-70 influenza immune response [29], suggesting that the role of ZAP-70 is duplicated in effective vaccines and is part of gaining vaccine immunity. Another study has investigated differential gene expression differences between older males and females from quadrivalent inactivated influenza virus vaccines [7]. Quadrivalent influenza vaccines can be incorporated into this framework to help understand how much of their differentially expressed genes are influenced by their vaccine type. Readers can validate our results by going to the VIGET website (https://viget.violinet.org/).

In the future, we plan to further examine how ontology combined with tools such as VIGET [16] can be used to further enhance our study of conditional vaccine immune responses given various experimental and clinical conditions. For example, by linking the VaximmutorDB data with the ImmPort data and Vaccine Ontology knowledge, we might be able to better understand the fundamental mechanisms of protective vaccine immune responses. Different ontology-based machine learning methods can also be explored to improve our prediction of new vaccine immune correlates and mechanisms.

5. Conclusions

In summary, the development of this ontology-based framework is able to aid data standardization and help guide further novel insights from the Omics data including the Omics metadata stored in the ImmPort database and Omics raw data from GEO. These mechanisms have been used to find differences between the pathways from different vaccines given different vaccine types and biological sexes. Moreover, this framework can also be expanded to other vaccines and vaccine categories.

6. Acknowledgements

This project was funded by a NIH-NIAID UH2 grant (1UH2AI132931-01A1) and a NIH-NIAID U24 grant (1U24AI171008). We appreciate the support provided by the ImmPort consortium. EO, YH and GW contributed to the overall study design; JZ, AMM and YH provided manual verification and ontology editing; EO, GW contributed to programming and data analysis; YH prepared the initial manuscript, AH prepared the updated manuscript for the paper and all authors contributed to the manuscript writing and reviews. All authors read and approved the final manuscript.

7. References

[1] **Standing up to infectious disease.** *Nat Microbiol.* 2019;**4**:1.

[2] Stutzer C, Richards SA, Ferreira M, Baron S, Maritz-Olivier C. **Metazoan Parasite Vaccines: Present Status and Future Prospects.** *Front Cell Infect Microbiol* 2018, **8**:67.

[3] Zhang C, Maruggi G, Shan H, Li J. **Advances in mRNA Vaccines for Infectious Diseases.** *Front Immunol* 2019, **10**:594.

[4] Berke K, Sun P, Ong E, Sanati N, Huffman A, Brunson T, Loney F, Ostrow J, Racz R, Zhao B *et al*: **VaximmutorDB: A Web-Based Vaccine Immune Factor Database and Its Application for Understanding Vaccine-Induced Immune Mechanisms**. *Front Immunol* 2021, **12**:639491.

[5] Beijnen EMS, Odumade OA, Haren SDV: **Molecular Determinants of the Early Life Immune Response to COVID-19 Infection and Immunization**. *Vaccines (Basel)* 2023, **11**(3).

[6] Zheng J, Li H, Liu Q, He Y: **The Ontology of Biological and Clinical Statistics (OBCS)-based statistical method standardization and meta-analysis of host responses to yellow fever vaccines**. *Quant Biol* 2017, **5**(4):291-301.

[7] Yang J, Huang X, Zhang J, Fan R, Zhao W, Han T, Duan K, Li X, Zeng P, Deng J *et al*: **Sex-specific differences of humoral immunity and transcriptome diversification in older adults vaccinated with inactivated quadrivalent influenza vaccines**. *Aging (Albany NY)* 2021, **13**(7):9801-9819.

[8] Hilleman MR, Carlson AJ, Jr., McLean AA, Vella PP, Weibel RE, Woodhour AF: **Streptococcus pneumoniae polysaccharide vaccine: age and dose responses, safety, persistence of antibody, revaccination, and simultaneous administration of pneumococcal and influenza vaccines**. *Rev Infect Dis* 1981, **3 Suppl**:S31-42.

[9] Bhattacharya S, Dunn P, Thomas CG, Smith B, Schaefer H, Chen J, Hu Z, Zalocusky KA, Shankar RD, Shen-Orr SS *et al*: **ImmPort, toward repurposing of open access immunological assay data for translational and clinical research**. *Scientific data* 2018, **5**:180015.

[10] Zheng J, Li H, Liu Q, He Y: **The Ontology of Biological and Clinical Statistics (OBCS)-based statistical method standardization and meta-analysis of host responses to yellow fever vaccines**. *Quantitative Biology* 2017, **5**(4):291-301.

[11] Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, He Y: **OBCS: The Ontology of Biological and Clinical Statistics**. In: *The 2014 International Conference on Biomedical Ontologies (ICBO 2014): 2014; Houston, TX, USA*. 2014: 1-6.

[12] Todd TE, Tibi O, Lin Y, Sayers S, Bronner DN, Xiang Z, He Y: **Meta-analysis of variables affecting mouse protection efficacy of whole organism Brucella vaccines and vaccine candidates**. *BMC bioinformatics* 2013, **14 Suppl 6**:S3.

[13] Ozgur A, Xiang Z, Radev DR, He Y: **Mining of vaccine-associated IFN-gamma gene interaction networks using the Vaccine Ontology**. *Journal of biomedical semantics* 2011, **2 Suppl 2**:S8.

[14] Ong E, Sun P, Berke K, Zheng J, Wu G, He Y: **VIO: ontology classification and study of vaccine responses given various experimental and analytical conditions**. *BMC Bioinformatics* 2019, **20**(Suppl 21):704.

[15] Zheng J, Harris MR, Masci AM, Lin Y, Hero A, Smith B, He Y: **The Ontology of Biological and Clinical Statistics (OBCS) for standardized and reproducible statistical analysis**. *J Biomed Semantics* 2016, **7**(1):53.

[16] Brunson T, Sanati N, Huffman A, Masci AM, Zheng J, Cooke MF, Conley P, He Y, Wu G: **VIGET: A web portal for study of vaccine-induced host responses based on Reactome pathways and ImmPort data**. *Front Immunol* 2023, **14**:1141030.

[17] Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: **OntoFox: web-based support for ontology reuse**. *BMC research notes* 2010, **3:175**:1-12.

[18] Storey, J. D., & Tibshirani, R. (2003). **Statistical significance for genomewide studies.** *Proc. Natl. Acad. Sci.* 2003, **100**(16), 9440–9445.

[19] He Y, Xiang Z, Zheng J, Lin Y, Overton JA, Ong E: **The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability**. *J Biomed Semantics* 2018, **9**(1):3.

[20] Xiang Z, Courtot M, Brinkman RR, Ruttenberg A, He Y: **OntoFox: web-based support for ontology reuse**. *BMC Res Notes* 2010, **3**:175.

[21] Xiang Z, Zheng J, Lin Y, He Y: **Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns**. *J Biomed Semantics* 2015, **6**:4.

[22] Overton JA, Vita R, Dunn P, Burel JG, Bukhari SAC, Cheung KH, Kleinstein SH, Diehl AD, Peters B: **Reporting and connecting cell type names and gating definitions through ontologies**. *BMC Bioinformatics* 2019, **20**(Suppl 5):182.

[23] Wu C, Hu Q, Ma D: **Development of an immune-related gene pairs signature for predicting clinical outcome in lung adenocarcinoma**. *Sci Rep* 2021, **11**(1):3611

[24] Guo L, Wu Q, Ma Z, Yuan M, Zhao S: **Identification of immune-related genes that predict prognosis and risk of bladder cancer: bioinformatics analysis of TCGA database**. *Aging (Albany NY)* 2021, **13**(15):19352-19374.

[25] Ciarambino T, Para O, Giordano M: **Immune system and COVID-19 by sex differences and age**. *Womens Health (Lond)* 2021, **17**:17455065211022262.

[26] Lacy P: **Mechanisms of degranulation in neutrophils**. *Allergy Asthma Clin Immunol* 2006, **2**(3):98-108.

[27] Eichelberger KR, Goldman WE: **Manipulating neutrophil degranulation as a bacterial virulence strategy**. *PLoS Pathog* 2020, **16**(12):e1009054.

[28] Petito E, Franco L, Falcinelli E, Guglielmini G, Conti C, Vaudo G, Paliani U, Becattini C, Mencacci A, Tondi F *et al*: **COVID-19 infection-associated platelet and neutrophil activation is blunted by previous anti-SARS-CoV-2 vaccination**. *Br J Haematol* 2023.

[29] Nguyen THO, Sant S, Bird NL, Grant EJ, Clemens EB, Koutsakos M, Valkenburg SA, Gras S, Lappas M, Jaworowski A *et al*: **Perturbed CD8(+) T cell immunity across universal influenza epitopes in the elderly**. *J Leukoc Biol* 2018, **103**(2):321-339.

Exploring the Use of Ontology Components for Distantly-Supervised Disease and Phenotype Named Entity Recognition

Sumyyah Toonsi[1,2,†], Şenay Kafkas[1,2,†] and Robert Hoehndorf[1,2,*]

[1] *Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

[2] *Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

Abstract

The lack of curated corpora is one of the major obstacles for Named Entity Recognition (NER). With the advancements in deep learning and development of robust language models, distant supervision utilizing weakly labelled data is often used to alleviate this problem. Previous approaches utilized weakly labeled corpora from Wikipedia or from the literature. However, to the best of our knowledge, none of them explored the use of the different ontology components for disease/phenotype NER under the distant supervision scheme. In this study, we explored whether different ontology components can be used to develop a distantly supervised disease/phenotype entity recognition model. We trained different models by considering ontology labels, synonyms, definitions, axioms and their combinations in addition to a model trained on literature. Results showed that content from the disease/phenotype ontologies can be exploited to develop a NER model performing at the state-of-the-art level. In particular, models that utilised both the ontology definitions and axioms showed competitive performance compared to the model trained on literature. This relieves the need of finding and annotating external corpora. Furthermore, models trained using ontology components made zero-shot predictions on the test datasets which were not observed by the models training on the literature based datasets.

Keywords

Named Entity Recognition, Text mining, ontologies

1. Introduction

Named Entity Recognition (NER) is a form of Natural Language processing (NLP) that aims to identify and classify named entities such as organisation, person, disease and genes in text. NER is a challenging task due to the nature of language which includes abbreviations, synonymous entities, and in general variable descriptions of entities.

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*] Corresponding author.

[†] These authors contributed equally.

sumyyah.toonsi@kaust.edu.sa (S. Toonsi); senay.kafkas@kaust.edu.sa (Ş. Kafkas); robert.hoendorf@kaust.edu.sa (R. Hoehndorf)

0000-0003-4746-4649 (S. Toonsi); 0000-0001-7509-5786 (Ş. Kafkas); 0000-0001-8149-5890 (R. Hoehndorf)

CEUR Workshop Proceedings (CEUR-WS.org)

Early methods for NER used dictionaries due to their applicability and time efficiency. Lexical approaches such as the NCBO (National Center for Biomedical Ontology) annotator [1], ZOOMA [2], and the OBO (Open Biological and Biomedical Ontologies) annotator [3] are not able to recognise new concepts and cannot detect all variations of expressions. This is because once dictionaries are constructed with terms, they can only find exact matches to those terms. Hence, dictionary-based approaches suffer from low recall.

With the emergence of machine learning, better NER methods were developed. This was possible through exposing statistical models to curated text where mentions of entities are identified by human curators and provided to these models. Subsequently, these models were able to generalize to unseen entities better than previous methods. For instance, GNormPlus [4] was developed to find gene/protein mentions using a supervised model which demonstrated competitive results at the time. Although supervised methods showed remarkable improvements in performance, they require curated instances for the model to learn. That is, the model expects instances of text where mentions of entities are clearly provided to learn to distinguish concepts of interest. This becomes a serious problem when one wants to recognise a novel/unexplored concept. Moreover, supervised methods often fail to recognise concepts uncovered by the curated corpora.

To alleviate the need for curated corpora, distant-supervision was explored for NER. In particular, distantly supervised models are trained on a weakly labeled training set, i.e., obtained from an imprecise source. For instance, dictionaries could be used to annotate text with exact matches which can produce both false positives and false negatives. Methods like BOND[5], PatNER[6], ChemNER[7], PhenoTagger [8], Conf-MPU [9] and Dong and colleagues [10] demonstrated the potential of distant supervision for NER. The aforementioned methods created weakly labeled sets using labels and synonyms found in ontologies/vocabularies to extract training instances from unlabeled corpora. Later, these instances were used to train different models which in some cases outperformed state-of-the-art methods.

Inspired by the advances achieved by distant supervision, we explored the contribution of different components of ontologies (Labels and synonyms, definitions, and complex axioms) to the task of NER under the distant supervision scheme. In all of the previously mentioned distantly-supervised NER methods, only labels and synonyms of ontologies/vocabularies were used to create the weakly labeled corpora from literature. The use of different ontology components to develop NER models has not been comprehensively explored for diseases/phenotypes. In addition to the use of labels and synonyms, in this study, we go a step further to explore the use of definitions and axioms to develop a disease/phenotype NER model. We hypothesize that the dense and rich knowledge found in ontologies can be used to develop NER models without the need of external corpora such as literature abstracts. We conducted our experiments on disease and phenotype entity recognition because, the study of diseases and phenotypes is important for understanding disease diagnosis, treatment and epidemiology.

2. Materials and Methods

2.1. Ontologies, literature resource and benchmark corpora

2.1.1. Ontologies

We used the Disease Ontology (DO) [11] on 15/April/2022) (downloaded on 1/March/2022) and the MEDIC vocabulary [12] in our study. DO is an ontology from the Open Biomedical Ontologies (OBO) [11], whereas MEDIC is a vocabulary of disease terms represented in the Web Ontology Language (OWL) [12]. We used the Human Phenotype Ontology (HPO) [13] (downloaded on 5/Jan/2022) for the phenotype concepts.

2.1.2. Literature

We used Medline [14] as a literature resource to generate our abstract-based weakly labeled dataset. To select abstracts that cover ontology concepts, we used an in-house index covering 32,923,095 Medline records (downloaded on Dec-15-2022) generated using Elasticsearch [15].

2.1.3. Benchmark corpora

To evaluate the named entity recognition models, we used four benchmark corpus; the NCBI–Disease Corpus [16] and the MedMentions Corpus (disease and phenotype) [17] and GSC+ [18]. NCBI–Disease is a widely used corpus where disease mentions are annotated and reviewed by multiple annotators. MedMentions is a large corpus annotated by an extensive set of Unified Medical Language System (UMLS) concepts. We selected the abstracts with disease annotations from MedMentions and named this the MedMentions–disease Corpus. To form this corpus, we used UMLS-to-MESH mappings from UMLS to obtain the MESH codes and selected the disease concepts which exist in our disease dictionary (described in section 2.2). Similarly, we selected the abstracts with phenotype concepts where we found mappings from UMLS-to-HPO and named this dataset as MedMentions–phenotypes. GSC+ is a widely used benchmarking dataset covering phenotype concepts particularly from HPO. We used the test dataset version released by [8]. Table 1 shows the distribution of the abstracts and annotations in the four benchmark corpora.

Table 1
Statistics of benchmark corpora

Corpus	Abstracts	Annotations
NCBI–disease train	593	5146
NCBI–disease dev	100	788
NCBI–disease test	100	960
MedMentions–disease test	879	3726
MedMentions–phenotype train	1291	6772
MedMentions–phenotype dev	428	2287
MedMentions–phenotype test	405	2190
GSC+ test	228	1933

2.2. Dictionary generation

We generated and used two dictionaries to weakly label Medline abstracts for disease and phenotype concepts. To generate our dictionaries, first, we extracted the labels and synonyms of all concepts from MEDIC, DO and HPO. Second, we filtered out the possible ambiguous labels/synonyms which are often stop words, short labels/synonyms (1 or 2 character long) and labels/synonyms shared by two different concepts from the dictionary. For example, DO contains a synonym which is "go" for the "geroderma osteodysplasticum" concept (DOID:0111266). The synonym "go" is ambiguous with the verb "go". Filtering out ambiguous names is a common practice used in text mining workflows that rely on lexical matches. We used the Natural Language Toolkit (NLTK) stop words [19] and filtered out any exact match with the labels/synonyms in MEDIC and DO and HPO. In both sources, we did not find any match with the list of stop words. We also filtered out the labels/synonyms having less than 3 characters to avoid false positives. Additionally, for the generation of the dictionary for diseases, we filtered out all the disease labels/synonyms which exactly match with protein labels/synonyms from the HUGO Gene Nomenclature Committee (HGNC) Database [20] to avoid false positive matches with protein names. Third, we generated the plural form of each label/synonym by using the Inflect Python module [21]. For example, the module generates “tetanic cataracts” for the given multi-word term, “tetanic cataract” (DOID:13822). Our final disease dictionary covers 244,903 disease labels and synonyms of 29,374 distinct concepts from MEDIC and DO. The final phenotype dictionary covers 79,010 phenotype labels and synonyms of 14,631 distinct concepts from HPO.

2.3. Ontology components used

An ontology O, as previously described in [22], has four main components:

- Classes and relations, where classes and relations are assigned unique identifiers.
- Domain vocabulary, where labels and synonyms are linked to ontology classes and relations.
- Textual definitions, where descriptions about classes and relations are provided, usually in natural language.
- Formal axioms, where relations between concepts are described in some formal language and possibly linked to other ontologies and sources.

We used labels and synonyms, textual definitions, and formal axioms components separately to create weakly labeled corpora and the statistics are reported in Table 2.

Table 2
Statistics of used ontology components

Component	DO and MEDIC	HPO
Labels/synonyms	35,333	16,307
Definitions	9,435 and 19,939 dummy	10,202 and 2,451 dummy
Axioms	30,834	37,062

2.4. Training dataset construction

2.4.1. Abstracts from literature

To generate the training set for distant supervision, first, we retrieved the relevant literature by searching the indexed Medline for the exact match of each label/synonym from the dictionaries. We retrieved the top [1-5] Medline abstracts/titles hits per concept that is identified based on the default Elastic Search Engine relevance scoring settings (TF-IDF [23] based scoring). Second, we used the dictionaries and annotated the downloaded abstracts lexically and converted the annotations to the I-O-B format (a common format for tagging tokens in a chunking task where B indicates the first token (Beginning) of an annotation, I subsequent (Inside) token of the same annotation and O representing a token that is not annotated (Outside)) [24] by using spaCy [25]. Finally, we obtained two sets of corpora; one for the disease concepts and the other for the phenotype concepts. We found 16,307 distinct phenotype labels/synonyms belonging to 6,962 classes from HPO in at least one Medline record by searching the indexed literature. These concepts are covered by 16096, 31372, 46032, 60098 and 74087 distinct Medline abstracts/titles at top 1, 2, 3, 4, 5 hits respectively, and we used them as our training sets for phenotypes. We found 35,333 distinct disease labels/synonyms linked to 8,400 distinct concepts from MEDIC and DO in at least one Medline records. These concepts are covered by 41698, 81007, 118295, 154060 and 187462 distinct Medline abstracts/titles at top 1, 2, 3, 4, 5 hits respectively and we used as our training sets for disease concepts.

Table 3
Example of using the class DOID:0040099 to create different weakly labeled sets. Text in bold refers to text annotated as B/I classes in the IOB format.

Component	Ontology representation	Dataset representation
Labels	name: Livedoid vasculitis	**Livedoid vasculitis**
Synonyms	synonym: “livedoid vasculopathy” EXACT	**Livedoid vasculopathy**
Axioms	DOID:0040099 SubClassOf DOID:865	**Livedoid vasculitis** is a **vasculitis**
Definitions	“A vasculitis with purpuric ulcers.”	A **vasculitis** with purpuric **ulcers**.

2.4.2. Labels and synonyms

Using the direct labels and synonyms from ontologies, we created two sets for phenotypes and diseases. For phenotypes, the labels and synonyms extracted from HPO were directly considered as positives as shown in Table 3. We used the labels and synonyms from DO and added MEDIC as well. The labels and synonyms were retrieved from the dictionary described in 2.2.

2.4.3. Definitions

Definitions in DO are available in natural language. To associate the concept with its definition, we added the concept label/synonyms to the beginning of a definition as shown in Table 3. For concepts which lacked definitions, we simply included their labels/synonyms with a dummy sentence replicated for all. For instance, if a disease X does not have a definition, its dummy definition is “X is a disease”. Since definitions can included other concepts (e.g. parent concepts) in their description, mentions of such concepts can be troublesome. To partially resolve this issue, we annotated the definitions with the dictionaries described in 2.2 Matches against the dictionaries were treated as positive mentions of concepts. In total, we retrieved 9,435 definitions from DO and used dummy definitions for 19,939 concepts. For phenotypes, we included definitions for 10,202 concepts and used dummy definitions for 2,451 concept.

2.4.4. Axioms

Axioms are not readily available for natural language tasks since they are expressed in formal language. To tackle this issue, we first processed axioms as previously described in [26]. Next, we replaced ontology identifiers with their labels/synonyms. We also included axioms which reference external ontologies and replaced their identifiers with names as shown in Table 3.

For diseases, we used 30,834 axioms from DO. For phenotypes, we included 37,062 axioms from HPO. Axioms of both concepts included references to external ontologies which we downloaded and processed to map their identifiers to their names. The external ontologies that were included are: the Basic Formal Ontology (BFO) [27], the Chemical Entities of Biological Interest (ChEBI) [28], the Cell Ontology (CL) [29], the Gene Ontology (GO), the Relation Ontology (RO) [30], and the Uber-anatomy Ontology (UBERON) [31].

2.5. Named entity recognition using distant supervision

NER refers to identifying boundaries of entity mentions in text (disease and phenotype mentions in our case). We used distant supervision to train our models by using BioBERT to recognise disease and phenotype mentions in text. Figure 1 depicts the system overview.

BioBERT is a BERT (Bidirectional Encoder Representations from Transformers) [32] pre-trained language model based on large biomedical corpora. BERT is a contextualized word representation model trained using masked language modeling. It provides self-supervised deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts. The pre-trained BERT model can be fine-tuned with an additional output layer to generate models for various desired NLP tasks. We used *simpletransformers* [33] which provides a wrapper model to distantly supervise an entity recognition model. More specifically, the wrapped model is used to fine-tune BERT models by adding a token-level classifier on top that classifies tokens into one of the output classes which are I-O-B (Inside-Outside-Beginning). In the training phase, our models are initialised with weights from BioBERT-Base v1.1 [34] and then fine-tuned on the disease and phenotype entity recognition task using our training corpora.

3. Results

We set up our experiments on four separate benchmarking corpora covering phenotype and disease concepts; NCBI–disease, MedMentions–disease, MedMentions–phenotype and GCS+. We reported our NER results using the Precision, Recall and F-score metrics. We used a relaxed scheme to calculate the metrics where we considered any partial overlap between the prediction and the curated annotations to be a true positive. That is, predictions are considered to be

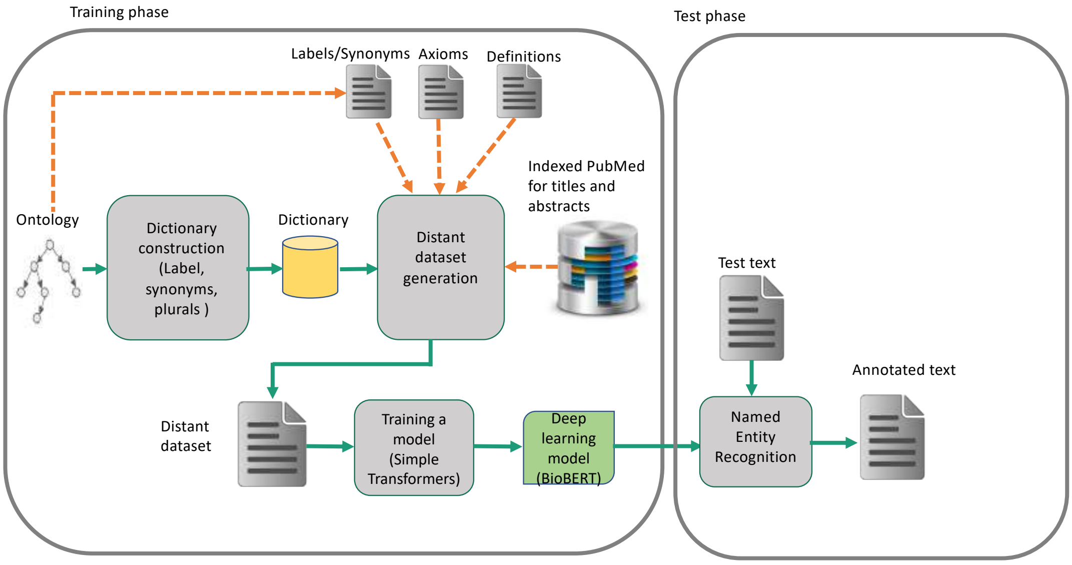

Training phase
Test phase
Labels/Synonyms
Axioms
Definitions
Ontology
Dictionary construction (Label, synonyms, plurals)
Dictionary
Distant dataset generation
Indexed PubMed for titles and abstracts
Test text
Annotated text
Distant dataset
Training a model (Simple Transformers)
Deep learning model (BioBERT)
Named Entity Recognition

Figure 1: System Overview
This figure depicts the training and test phases in our system. In the training phase, we used ontologies to create a dictionary from the labels, synonyms and their plural forms. We used this dictionary to create distant datasets from Medline abstracts and different ontology parts (labels/synonyms, axioms and definitions). Later, this distant dataset is used for training a BioBERT NER model by using the SimpleTranformers wrapper. In the test phase, the trained model is tested on different benchmarking corpora.

positives whenever the indices (locations in text) of the prediction and the curated annotations overlap.

Table 4 shows the performance of the disease NER models which are distantly supervised on different ontology components or on abstracts (best F1-score is achieved at top 1, see Additional File 1) on the disease test sets (see Table 1). For the sake of comparison, we also included a supervised BioBERT model that is trained on the NCBI-disease training set. Our results showed that supervised BioBERT trained on the curated set performed the best on NCBI–disease (0.94 F1-score) because concepts are highly conserved in this dataset. To fairly compare the performance of the methods, we further evaluated the models on the MedMentions–disease dataset. Results showed that the distantly supervised models (trained on abstracts and definitions plus axioms) achieved higher F1 scores (0.68 for abstracts and 0.67 for definitions and axioms) compared to the model trained on the curated set (0.66 F1-score) which is actually biased towards the NCBI–disease dataset (we found out there is 80% overlap in concept IDs between NCBI training and test sets). The models trained on solely labels and synonyms, axioms, definitions showed lower F1-score compared to the model trained on abstracts. On the other hand, the model trained on definitions plus axioms achieved a competitive F1-score compared to the model trained on abstracts. This result is more evident on the MedMentions-disease test set.

Table 4
Disease NER results

Corpus	Precision	Recall	F1
NCBI-disease			
Labels and synonyms	0.64	0.36	0.46
Axioms	0.68	0.59	0.63
Definitions	0.87	0.80	0.83
Definitions and axioms	0.91	0.76	0.83
Literature abstracts	0.92	0.81	0.86
Curated NCBI train	0.91	0.96	0.94
MedMentions-disease			
Labels and synonyms	0.41	0.26	0.32
Axioms	0.43	0.42	0.43
Definitions	0.48	0.82	0.61
Definitions and axioms	0.58	0.79	0.67
Literature abstracts	0.60	0.78	0.68
Curated NCBI train	0.58	0.77	0.66

Table 5
Phenotype NER Results

Corpus	Precision	Recall	F1
MedMentions-phenotype			
Labels and synonyms	0.33	0.75	0.46
Axioms	0.31	0.58	0.40
Definitions	0.47	0.80	0.59
Definitions and axioms	0.55	0.77	0.64
Literature abstracts	0.60	0.82	0.69
Curated MedMentions train	0.61	0.79	0.69
GSC+			
Labels and synonyms	0.32	0.71	0.44
Axioms	0.40	0.60	0.48
Definitions	0.61	0.77	0.68
Definitions and axioms	0.65	0.74	0.69
Literature abstracts	0.73	0.78	0.75
Curated MedMentions train	0.61	0.53	0.57

Table 5 presents the performance of the models in phenotype NER on the GSC+ and MedMentions-phenotype test datasets. We included the MedMentions-phenotype dataset to thoroughly test our models and to train the supervised model on sufficient data. With the inclusion of context at a large scale, the model trained on the weakly labelled abstracts achieved the highest F1-score (0.69 F1-score on MedMentions-phenotype and 0.75 on GSC+) compared to other models. On the other hand, the model trained on the curated set was not robust to the change of dataset as it performed poorly on GSC+ (0.57 F1-score). We observed 6% discrepancy between the model trained on abstracts and the model trained on weakly labelled definitions plus axioms. We discuss the reasons of this discrepancy in detail in the "Discussion" section.

4. Discussion

Our main goal was to explore whether ontology components can help to develop distantly supervised disease/phenotype entity recognition models which are competitive to the state-of-the-art. To that end, we exploited ontological components to create textual context using the labels/synonyms, axioms and definitions. We observed that utilising the context in ontologies via distant supervision aids in developing a NER model at the state-of-the-art level. While the models trained solely on labels and synonyms achieves lowest simply due to lack of context; the models incorporating context such as axioms and definitions improved the performance upon the models that lack context.

The disease NER model trained on the axioms and definitions achieved competitive F1-score compared to the model trained on the abstracts only. However, we observed 6% discrepancy between the phenotype NER models trained on the abstracts (best F1-score is achieved at top 2) and axioms and definitions together. To investigate the reason for this discrepancy, we focused on the False Positive (FP) predictions that we achieved on the GSC+ test corpus. The model trained on the weakly labeled abstracts produced 440 FPs while the model trained on the phenotype definitions and axioms produced 608 FPs. We found that 184 out of 608 FPs are produced distinctly by the model trained on definitions and axioms and not by the one trained on the abstracts. We randomly sampled 20 FPs from these 184 FPs for further manual analysis. Our manual analysis on these 20 FPs showed that all of them were actually True Positives but have been missed by the GSC+ dataset. For example, we found "Uniparental disomy" (HP:0032382) in PMID:8103288 was captured correctly by the model but was missed by GSC+ annotations. More importantly, we observed that the majority of the FPs were not introduced in the definitions and axioms training corpus but were rather predicted as zero-shot instances (i.e. instances that were not seen by the model during training). For example, "Angelman syndrome" in PMID:8786067 which does not correspond to any label/synonyms in HPO and does not exist in the corpus was annotated by the model trained on definitions and axioms. Furthermore, the model trained on literature abstracts did not have these FPs since they were specifically included as O classes in the training set. Details on our manual analysis can be found in the Additional Files 1.

We conducted our study on DO and HPO. These ontologies are widely used and therefore contain dense content which can help to generate sufficiently large weakly label datasets. Although the approach is generic and its utility can be explored for any given ontology; the performance would depend on the density of the content of the ontology of choice. That is, if the ontology does not sufficiently describe a concept, it is not possible to obtain a well-performing model.

5. Conclusion

In conclusion, our analysis showed that the ontology components can provide a suitable corpus to build a NER model that is competitive to state-of-the-art. This alleviates the need for annotating a large number of abstracts and facilitates the creation of weakly labeled training corpora. Easily obtained corpora are desirable since they reduce both the computational and time overheads. To our best knowledge, this is the first work that uses ontology axioms to build disease/phenotypes NER models.

Additionally, the models trained on ontology components were capable of zero-shot learning on the test datasets. This was not the cases for the models trained on curated sets and the models trained on the large weakly labeled literature abstracts. Our approach is generic and its utility can be explored with any other given ontology which has sufficient content that describes the concept of interest.

Acknowledgments

We thank Dr. Mahmut Uludağ for his technical assistance in processing MEDLINE data.

This work has been supported by funding from King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. URF/1/4355-01-01, URF/1/4675-01-01, URF/1/4697-01-01, URF/1/5041-01-01, REI/1/5334-01-01, FCC/1/1976-46-01 and FCC/1/1976-34-01.

References

[1] C. Jonquet, N. H. Shah, M. A. Musen, The open biomedical annotator, in: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09, San Francisco, CA, USA, 2009, pp. 56–60.

[2] M. Kapushesky, et al., Gene expression atlas update–a value-added database of microarray and sequencing-based functional genomics experiments, Nucleic Acids Research 40 (2011) D1077–D1081. URL: https://doi.org/10.1093/nar/gkr913. doi:10.1093/nar/gkr913.

[3] M. Taboada, H. Rodriguez, D. Martinez, M. Pardo, M. J. Sobrido, Automated semantic annotation of rare disease cases: a case study, Database 2014 (2014) bau045–bau045. URL: https://doi.org/10.1093/database/bau045. doi:10.1093/database/bau045.

[4] C.-H. Wei, H.-Y. Kao, Z. Lu, GNormPlus: An integrative approach for tagging genes, gene families, and protein domains, BioMed Research International 2015 (2015) 1–7. URL: https://doi.org/10.1155/2015/918710. doi:10.1155/2015/918710.

[5] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, C. Zhang, Bond: Bert-assisted open-domain named entity recognition with distant supervision, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1054–1064. URL: https://doi.org/10.1145/3394486.3403149. doi:10.1145/3394486.3403149.

[6] X. Wang, Y. Guan, Y. Zhang, Q. Li, J. Han, Pattern-enhanced named entity recognition with distant supervision, in: 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 818–827. doi:10.1109/BigData50022.2020.9378052.

[7] X. Wang, V. Hu, X. Song, S. Garg, J. Xiao, J. Han, ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp.

5227–5240. URL: https://aclanthology.org/2021.emnlp-main.424. doi:10.18653/v1/2021.emnlp-main.424.

[8] L. Luo, S. Yan, P.-T. Lai, D. Veltri, A. Oler, S. Xirasagar, R. Ghosh, M. Similuk, P. N. Robinson, Z. Lu, PhenoTagger: a hybrid method for phenotype concept recognition using human phenotype ontology, Bioinformatics 37 (2021) 1884–1890. URL: https://doi.org/10.1093/bioinformatics/btab019. doi:10.1093/bioinformatics/btab019.

[9] K. Zhou, Y. Li, Q. Li, Distantly supervised named entity recognition via confidence-based multi-class positive and unlabeled learning, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7198–7211. URL: https://aclanthology.org/2022.acl-long.498. doi:10.18653/v1/2022.acl-long.498.

[10] H. Dong, V. Suárez-Paniagua, H. Zhang, M. Wang, A. Casey, E. Davidson, J. Chen, B. Alex, W. Whiteley, H. Wu, Ontology-driven and weakly supervised rare disease identification from clinical notes, BMC Medical Informatics and Decision Making 23 (2023). URL: https://doi.org/10.1186/s12911-023-02181-9. doi:10.1186/s12911-023-02181-9.

[11] L. M. Schriml, et al., Human Disease Ontology 2018 update: classification, content and workflow expansion, Nucleic Acids Research 47 (2018) D955–D962. URL: https://doi.org/10.1093/nar/gky1032. doi:10.1093/nar/gky1032.

[12] A. P. Davis, T. C. Wiegers, M. C. Rosenstein, C. J. Mattingly, MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database, Database 2012 (2012). URL: https://doi.org/10.1093/database/bar065. doi:10.1093/database/bar065, bar065.

[13] S. Köhler, et al., Expansion of the human phenotype ontology (HPO) knowledge base and resources, Nucleic Acids Research 47 (2018) D1018–D1027. URL: https://doi.org/10.1093/nar/gky1105. doi:10.1093/nar/gky1105.

[14] NCBI, Pubmed, 1996. https://pubmed.ncbi.nlm.nih.gov/, Last accessed on 2022-04-18.

[15] N. Elastic, Swiftype, Elastic search, 2010. https://www.elastic.co/, Last accessed on 2022-04-18.

[16] R. I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: A resource for disease name recognition and concept normalization, Journal of Biomedical Informatics 47 (2014) 1–10. URL: https://doi.org/10.1016/j.jbi.2013.12.006. doi:10.1016/j.jbi.2013.12.006.

[17] S. Mohan, D. Li, Medmentions: A large biomedical corpus annotated with umls concepts, 2019. URL: https://arxiv.org/abs/1902.09476. doi:10.48550/ARXIV.1902.09476.

[18] M. Lobo, A. Lamurias, F. M. Couto, Identifying human phenotype terms by combining machine learning and validation rules, BioMed Research International 2017 (2017) 1–8. URL: https://doi.org/10.1155/2017/8565739. doi:10.1155/2017/8565739.

[19] I. Brigadir, Nltk stop words, 2019. https://github.com/igorbrigadir/stopwords/blob/master/en/nltk.txt, Last accessed on 2022-09-14.

[20] S. Tweedie, B. Braschi, K. Gray, T. E. M. Jones, R. L. Seal, B. Yates, E. A. Bruford, Genenames.org: the HGNC and VGNC resources in 2021, Nucleic Acids Research 49 (2020) D939–D946. URL: https://doi.org/10.1093/nar/gkaa980. doi:10.1093/nar/gkaa980.

[21] P. Dyson, Inflect python module, 2022. https://pypi.org/project/inflect/, Last accessed on 2022-09-14.

[22] R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, The role of ontologies in biological and biomedical research: a functional perspective, Briefings in bioinformatics 16 (2015) 1069–

1080.

[23] C. Sammut, G. I. Webb (Eds.), TF–IDF, Springer US, Boston, MA, 2010, pp. 986–987. URL: https://doi.org/10.1007/978-0-387-30164-8_832. doi:10.1007/978-0-387-30164-8_832.

[24] L. A. Ramshaw, M. P. Marcus, Text chunking using transformation-based learning, in: ACL Third Workshop on Very Large Corpora, 1995, pp. 82–94. doi:https://doi.org/10.48550/arXiv.cmp-lg/9505040.

[25] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017. To appear.

[26] F. Z. Smaili, X. Gao, R. Hoehndorf, Onto2Vec: joint vector-based representation of biological entities and their ontology-based annotations, Bioinformatics 34 (2018) i52–i60. URL: https://doi.org/10.1093/bioinformatics/bty259. doi:10.1093/bioinformatics/bty259.

[27] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, The MIT Press, Cambridge, Massachusetts;London, England;, 2015;2016;.

[28] J. Hastings, et al., Chebi in 2016: Improved services and an expanding collection of metabolites, Nucleic acids research 44 (2016) D1214—9. URL: https://europepmc.org/articles/PMC4702775. doi:10.1093/nar/gkv1031.

[29] T. Bakken, L. Cowell, B. D. Aevermann, M. Novotny, R. Hodge, J. A. Miller, A. Lee, I. Chang, J. McCorrison, B. Pulendran, et al., Cell type discovery and representation in the era of high-content single cell phenotyping, BMC bioinformatics 18 (2017) 7–16.

[30] R. P. Huntley, M. A. Harris, Y. Alam-Faruque, J. A. Blake, S. Carbon, H. Dietze, E. C. Dimmer, R. E. Foulger, D. P. Hill, V. K. Khodiyar, et al., A method for increasing expressivity of gene ontology annotations using a compositional approach, BMC bioinformatics 15 (2014) 1–11.

[31] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, M. A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome biology 13 (2012) 1–20.

[32] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, in: Proceedings of the 2019 Conference of the North, Association for Computational Linguistics, 2019. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[33] T. C. Rajapakse, Simple transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.

[34] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert github respository, 2019. (https://github.com/dmis-lab/biobert).

A. Appendix

- Additional file 1 — AdditionalFile1.xls First sheet name as "performance_on_abstracts" contains the performances of the models trained on the weakly labeled abstract datasets selected based on top [1-5] hits from the ElasticSearch Index. Second sheet named as "manual_error_analysis" contains our manual analysis results on the False Positives from the GSC+ dataset. The file is available from github: https://github.com/bio-ontology-research-group/OntoNER

An Extendible Realism-Based Ontology for Kinship

Michael Rabenberg [1], Anuwat Pengput [1] and Werner Ceusters [1]

[1] *University at Buffalo, 77 Goodell Street, Buffalo NY, 14203, USA*

Abstract

Adequately representing kinship relations is crucial for a variety of medical and biomedical applications. Several kinship ontologies have been proposed but none of them have been designed thus far in line with the Basic Formal Ontology. In this paper, we propose a novel kinship ontology that exhibits the following characteristics: (1) it is fully axiomatized in First Order Logic following the rules governing predicate formation as proposed in BFO2020-FOL, (2) it is modularized in 6 separate files written in the Common Logic Interface Format (CLIF) each one of which can be imported based on specific needs, (3) it provides bridging axioms to and from SNOMED CT, and (4) it contains an extra module with axioms which would not be literally true when phrased naively but are crafted in such a way that they highlight the unusual kinship relations they represent and can be used to generate alerts on possible data entry mistakes. We describe design considerations and challenges encountered.

Keywords

Kinship ontology, SNOMED CT, BFO2020

1. Introduction

Ontologies are rarely developed completely from scratch these days, even if they don't make use of any of the few upper ontologies that are around. Re-using ontologies, primarily parts thereof, has even become standard practice, specifically for ontologies rendered in OWL. However, it is not free of risks. Several issues with re-used content have been reported in BioPortal ontologies such as duplicated classes and object properties, inconsistent utilization of reused properties and redundant class hierarchies [1]. Tools such as ROBOT can detect certain syntactic errors in source ontologies used for import [2], and reasoners can detect logical errors within the boundaries of the logic they are designed for. Neither, however, can prevent the most common types of representation errors. One such type is caused by '*using OWL just as a syntax and ignoring its open-world semantics*' [3, p11]. That is for instance the case with the majority of assertions of the form '*disease - has symptom - some - symptom*' in the Disease Ontology [4]: the OWL semantics requires such statement to be true for ALL instances of the named disease, which is rarely the case. Although some recommend competency question-driven ontology authoring as a means to test the internal quality of an ontology before its release [5], doing so may result in '*tweaking the axioms so that the reasoner "gets the right answer" ignoring what else the axioms might entail*' [3, p11]. Another form of tweaking is simply removing the axioms that otherwise would come with imported classes or object-properties, or creating axiom-less classes and object-properties with vague textual definitions and names that are slightly altered from what is found in reference ontologies.

Unfortunately, it is not only representation errors that hamper ontology reuse, but also differing perspectives in otherwise similar domains, as well as the contexts and applications for which they have been developed. Simply importing terms and axioms from domain- and application ontologies that are claimed to be internally coherent and consistent – even logically – is therefore not enough. In many cases, a thorough manual inspection on top of semi-automated procedures is therefore required. In this

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: rabenbergm@gmail.com (A. 1); anuwatpe@buffalo.edu (A. 2); wceusters@gmail.com (A. 3)
ORCID: 0000-0001-9004-5322 (A. 1); 0000-0002-0273-1531 (A. 2); 0000-0002-2676-8689 (A. 3)

paper, we demonstrate the dilemmas that had to be faced in developing a realism-based kinship ontology for a specific use case while trying to maximally reuse what has already been developed. In Section 2, we explain realism-based ontology. In Section 3, we describe a use-case for our ontology, from which some of the relations in our ontology are derived. Section 4 contains a review of several extant kinship ontologies, including the one from which ours is primarily adapted. We describe our methodology in Section 5 and detail the results of our project in Section 6. Further discussion, including some of the more important design choices we made, appears in Section 7. Section 8 contains concluding remarks.

2. Realism-based ontology design

For something to be a realism-based ontology is for it (1) to be a representation of what is generically the case for a plurality of entities in reality, and (2) to be built out of smaller representations each one of which must be faithful to reality. If such a representation is expressed in terms of one or more representation languages, then at least two requirements must be met for all assertions that are part of the representation: (1) the symbols intended to denote entities denote only entities that exist or have existed and (2) the symbols which express relationships between these entities do so equally veridically, i.e. all posited relationships must obtain in reality. This means that the ontology must have a very precisely defined ontological commitment that is anchored in an ontological theory based on one or other form of realism. A requirement for high-quality ontologies – whether realism-based or not – is that the veridicality of all assertions made therein is verifiable. When part of the representation is expressed in a logical language, then the logical coherence and consistency of that part can be checked algorithmically. This is because logical connectors and quantifiers, if any at all, of such languages are precisely defined, as well as how they may be combined and what sorts of operations and transformations can be applied to them. Both of these requirements are satisfied by the Basic Formal Ontology (BFO): it is a domain-independent upper ontology [6] which is based on Ontological Realism [7]. Its most recent version, BFO2020-FOL, has recently been accepted as an ISO standard [8]. This includes an axiomatization in First Order Logic (FOL) for a fair part of its underlying philosophical principles and theories, enough to allow for sound spatial and temporal reasoning on top of the mere classification as offered by description logics.

3. Use case: history of cholangiocarcinoma in 'relatives'

Cholangiocarcinoma (CCA) is a significant public health problem in Thailand, especially in the northeastern region [9, 10]. CCA is a lethal bile duct cancer which in countries of Southeast Asia is associated with *Opisthorchis viverrini* (OV) infection [11]. Roughly 5,000 new cases of CCA are diagnosed annually, and at least 8 million people are infected with OV in Thailand. In 2013, the Khon Kaen University (KKU) developed a prospective cohort study called the Cholangiocarcinoma Screening and Care Program (CASCAP) to eliminate OV and CCA. This led in collaboration with the National Health Security Office and the Ministry of Public Health to a national policy to improve diagnosis and treatment for CCA, covering all primary, secondary, and tertiary cares [12]. Subjects enter the CASCAP program in one of two ways [13]. One is through screening performed in high-risk areas on the basis of voluntary enrollment. This includes a structured interview followed by an ultrasound screening for CCA. Patients suspected of having a CCA may obtain a confirmation diagnosis through Computerized Tomography (CT) or magnetic resonance imaging (MRI) [14]. Subjects can also enter when they are diagnosed as having a CCA in hospitals from the CASCAP network.

The CASCAP administration maintains a data repository about subjects exhibiting the following inclusion criteria: (1) living in northeastern Thailand, (2) being at least 40 years old, and (3) either of the following: (3a) ever having been infected with or treated for liver fluke, or (3b) ever having consumed raw freshwater fish with scales. Data in the repository is collected by means of six forms [15]. One of them is the Demographic Information and Enrollment form CCA-01, which researchers and public health officers use to register participants and collect from them demographic information as well as certain risk factors for CCA, amongst which is having a family history of CCA [16, 17]. It is the latter topic that inspired us as a demonstration use case for ontology re-use and adaptation in line with realism-based principles. Family history is in the CCA-01 form determined by means of a yes/no response to the question '*Do you have any relatives diagnosed with cholangiocarcinoma?*'. When this question is positively answered, the following options are offered as categories: (1) paternal grandfather or mother, (2) maternal grandfather or mother, (3) older aunt or uncle, (4) younger aunt or uncle, (5) father or mother, (6) son or daughter, (7) brother or sister, (8) nephew or niece, and (9) spouse. In this paper, we discuss the development methodology of our kinship ontology as well as the challenges encountered. In a companion paper, we elaborate on how to use it for quality control of the CASCAP data-repository.

4. Existing kinship ontologies

Chui et al. propose a kinship ontology that they call '$T_{kinship}$' [18]. The ontology is axiomatized in First Order Logic (FOL) and consists of 13 axioms. Axioms (1)–(8) are concerned with *natural-ancestor-of,* (9)–(12) with *spouse-of*, and axiom 13 with both. Chui et al. use the phrase 'ancestor' rather than the more specific term 'natural ancestor' and in general do not append 'natural' to terms that admit of 'non-natural' readings (e.g., 'grandparent'), but it is clear that when they use such terms they mean them to carry their blood-relative senses. Chui et al. also elaborate on how they think their axioms can be exploited so as to define some additional familial relations, such as the *has-natural-grandparent* relation.

Stevens et al. propose a different kinship ontology, the Family History Knowledge Database (FHKD), written in OWL 2 DL [19]. The FHKD was designed as a way of demonstrating OWL 2's features and testing automated reasoners. As they admit, some of the axioms in their ontology, if interpreted as genuine claims about reality, are seriously questionable; for example, some of their axioms concerning *siblinghood* imply that a given person, S, is sibling of S [19, p5]. But in light of the educational and testing purposes for which the FHKD was designed, it seems charitable to interpret such axioms not as genuine claims about reality. The work also demonstrates that even the expressive DL SROIQ(D) is not expressive enough to represent kinship.

KIN, a DL kinship ontology produced as a part of the Global Alliance for Genomics and Health Pedigree Standard project [20], contains a fairly small number of defined classes, with *person* and *sex* subsuming the rest, and a larger hierarchy of object properties, at the top of which are hasSex and isRelativeOf, and further down relations such as isGestationalCarrierOf and isMitochondrialDonorOf. KIN is said to allow using an OWL reasoner to automatically validate a family history graph and infer new relations, while the expansiveness of its object properties would allow for detailed descriptions and inferences concerning individuals. However, like FHKD, KIN contains some axioms that yield implausible results if interpreted as genuine claims about reality. For example, KIN holds that isRelativeOf is symmetric and transitive, thus implying (implausibly) that a given person, S, is relative of S. Furthermore, KIN's sex class-hierarchy is a bit peculiar. In addition to Female and Male, KIN contains OtherSex and UnknownSex, but OtherSex is a *subclass* of *Female or Male* (implying that it is *not* another sex, as might be had by some organism of an imaginable sexually ternary species); and KIN specifies that OtherSex is meant to cover cases in which '*It is not possible to accurately assess the applicability of male or female*' (making one wonder what its difference from UnknownSex is supposed to be) [21].

Cantone et al. present several DL kinship ontologies, some for SROIQ and one for EL++ [22]. They explicitly note (as mentioned above) that FHKB treats isSiblingOf as reflexive, and they appear to treat avoidance of such results as a side-constraint on the development of an acceptable kinship ontology. Among their goals, however, are to show some of the limitations of SROIQ and EL++, by showing that the SROIQ and EL++ ontologies that they consider are logically weaker than a different plausible set of kinship axioms that they call 'K_L.' For example, K_L contains an axiom to the following effect: x is relative of y iff (a) x and y are non-identical and (b) there is a sequence x…y every member of which with an immediate successor is relative of its immediate successor. As Cantone et al. point out, this axiom allows K_L to yield inferences that are unavailable to SROIQ and EL++ ontologies.

Although not itself a kinship ontology, SNOMED CT has a large number of kinship concepts, which are not classified as relations – in SNOMED CT called 'attributes' – but as 'person', a subhierarchy of 'social entity'. In what follows, we reference SNOMED CT concepts through the concatenation of 'sct_', the fully specified name (with the first letter of the fully specified name capitalized as in

SNOMED CT itself and spaces replaced by hyphens), a second underscore, and finally the concept's semantic tag. Two broad classes of kinship concepts are worth noting. First, there is the concept sct_Blood-relative_person and all the concepts falling under it. These concepts, unsurprisingly, specifically correspond to *blood-relative* categories of which individuals can be members. Examples include, in addition to sct_Blood-relative_person itself, sct_Natural-sibling_person and sct_Natural-child_person. Second, there are familial concepts in SNOMED CT that do *not* specifically correspond to blood-relative categories. Examples include sct_Niece_person and sct_Maternal-grandparent_person; one can be a niece or maternal grandparent of someone else without being a blood relative of that person.

5. Methodology

We built our ontology following a number of steps, thereby reiterating over previous steps whenever deemed necessary. These steps were taken to satisfy the following requirements for our ontology: (1) maximally re-use what is available and be maximally re-usable itself, (2) be able to represent all kinship relations required for the CCA-01 form, (3) be fully BFO2020-compatible and (4) fit in the logical framework set up to combine realism-based ontologies with concept-based ontologies such as SNOMED CT [23].

Step 1 consisted of manually inspecting the axioms in the existing kinship ontologies to assess the degree to which they can be read literally, i.e. the extent to which they are faithful to the aspects of reality to which they pertain. For example, consider axiom (A1) from the $T_{kinship}$ ontology:

(A1) $\forall x(\neg \text{ancestorOf}(x,x))$

If we thought there were counterexamples to this axiom – involving backwards time-travel and causal loops or some other such exotic phenomena – then we would not import it in our ontology. This step included the identification of axioms that are most often satisfied, yet not in general. An example is axiom (A2) as found in $T_{kinship}$.

(A2)
$$\begin{aligned}\forall x \forall y \quad & (\text{hasChild}(x,y) \\ & \equiv (\text{ancestorOf}(x,y) \land \\ & \quad \neg(\exists z\, (\text{ancestorOf}(x, z) \land \\ & \qquad \text{ancestorOf}(z,y)))))\end{aligned}$$

The sort of situation that counterexamples (A2) is indeed highly unusual: people do not ordinarily have natural children who are natural descendants of their own natural descendants. Readers familiar with the classic 1974 film *Chinatown* might recall that one of the big revelations toward the end of the movie is that one of the characters had fathered a child with his own daughter. This would be a situation of the sort at issue. Such axioms were however not excluded from our ontology but turned into related axioms in such a way as to mark the peculiarity of the situation.

Step 2 was to rewrite the accepted axioms so as to be fully compatible with the BFO2020-FOL axiomatization. Several transformations were to be considered. One was to rewrite predicates which for BFO are considered '*fantologically conceived*' [24]. Examples appear in axiom (A3) from $T_{kinship}$.

(A3)
$$\begin{aligned}\forall x \forall y\, & (\text{ancestorOf}(x,y) \\ & \rightarrow (\text{person}(x)\ \&\ \text{person}(y)))\end{aligned}$$

In FOL as used in $T_{kinship}$, 'ancestorOf' and 'person' are relations – binary and unary resp. – that hold for certain individuals in the domain of discourse. FOL allows one to predicate something about individuals in its domain without being bothered by any ontological commitment. While BFO might commit to 'ancestorOf' representing an instance-level formal relation [25], it would not commit to 'person' representing a formal relation but rather a universal instantiated by a particular at a time. A BFO-compatible FOL translation of (A3) would therefore be:

(A4) $\forall x \forall y$ (ancestorOf(x,y)
$\rightarrow$ ($\exists$t1$\exists$t2((instance-of(x, person, t1) &
instance-of(y, person, t2)))))

In general, any axiom in a source ontology that describes an individual as timelessly standing in some unary relation (as in FOL '*fantologically conceived*' [24]) or as timelessly being a member of a class (as in OWL) required revision for BFO2020-compatibility during Step 2. This is because under a realism-based perspective, such an individual is very likely a particular which instantiates at a time a universal. Such revision was also needed for cases expressing some individual's timelessly standing in some relation to some class member when such a relation would require time-indexing as per BFO's ontological commitment.

Step 3 consisted of replacing terms for relations and universals with terms that express better what is intended. Examples included replacing occurrences of 'ancestorOf' with occurrences of 'natural-ancestor-of' to make explicit that our axioms containing this expression are concerned with natural ancestry and not with some broader notion of ancestry, and replacing occurrences of 'person' with occurrences of 'human-being'.

Step 4 consisted in formulating axioms detailing relations between the kinship relations referenced on the CCA-01 form to the kinship relations completed thus far. Only one relation referenced on the form, the *has-spouse* relation, we did not address at this stage, because it already appeared in our collection. The other relations referenced on the form are all blood relations, despite the fact that the English version of this form confusingly contains some non–blood-relation-specific terms such as 'son', 'daughter', and 'sibling'; the Thai terms carry specifically blood-relation senses only.

We devised in Step 5 bridging axioms from the relations and universals referenced in the axioms produced thus far to corresponding kinship terms found in SNOMED CT and vice versa, as exemplified by (A6) and (A5) respectively. Although many of these axiom pairs can be written as biconditionals, we refrained from doing so for reason of modularization towards re-usability on a needs basis.

(A5) $\forall$x (individual-of(x, sct_Natural-child_person)
$\rightarrow$ ($\exists$y (has-natural-child(y, x))))

(A6) $\forall$x (($\exists$y (has-natural-child(y, x))
$\rightarrow$ individual-of(x, sct_Natural-child_person)))

In Step 6, we devised a set of axioms specifying certain highly unusual ancestry and spousal situations as unusual. These axioms are inspired by axioms from $T_{kinship}$ that we have rejected because they admit of counterexamples. An example is the rejected axiom (A2) (discussed above), from which we derived (A7).

(A7) $\forall$x$\forall$y$\forall$z((has-natural-child(x,y)
& natural-ancestor-of(x,z)
& natural-ancestor-of(z,y))
$\rightarrow$ occupy-unusual-ancestry-situation(x,y,z))

We note that we do not mean by 'unusual' 'counterintuitive'. For example, a spousehood relation between first cousins is not, in our view, counterintuitive, though we mark such a relation as unusual. A better gloss on 'unusual' is 'atypical'. Spousehood relations between first cousins are atypical—they just don't happen very often. Furthermore, they happen sufficiently infrequently that an unusualness axiom pertaining to them seems to serve valuable data-entry and -inspection purposes, as we explain in Section 6. We also emphasize that 'unusual', as we use it, is not meant to carry any normative weight. In calling a relation 'unusual', we do not mean to imply that it is bad, that it ought to be illegal, or that any other such normative fact obtains.

As a last step, we translated all axioms in CLIF following the schema of the BFO2020-FOL axiomatization in CLIF, used a parser-generator to transform the axiom collection in a Kowalski-rule base, and the latter as input for a reasoner for satisfiability testing [26]. Kowalski rules are a further transformation of FOL axioms after they have been translated into clausal normal form. Kowalski rules are logical implications of which the antecedent is formed by conjoining the atoms of the negative literals in a clause, and the consequent from the disjunction of the positive literals [27].

6. Results

Our kinship ontology consists of six modules. Some modules contain axioms whose definientia refer to relations or entities defined in other modules. However, some modules can be ignored when irrelevant for certain applications. That is for instance the case for the bridging axioms to and from SNOMED CT when SNOMED CT is not used in an application the ontology intends to serve.

Each axiom comes with a short textual description terminated by an index which is unique within and across all modules. This index can be used to import individual axioms as well as to create additional documentation containing detailed textual definitions and elucidations. It can also be used to link the ontology to a terminology.

The core module, ancestry.clif, consists of 56 axioms, each one of which belongs to one of the following categories: (1) replacements of the $T_{kinship}$ axioms, subindexed 'tkr'; (2) axioms for ordinary kinship relations requested on the CCA-01 form, subindexed 'cca'; (3) a recursive definition of the natural-ancestor-of relation, split into two axioms each subindexed 'nao'; (4) additional axioms for ordinary kinship relations, subindexed 'ak'; and (5) axioms linking universals referenced in this module to BFO categories, subindexed 'u'. Most of these relations are fairly 'ordinary' ones, such as the *natural-maternal-grandparent-of* relation and the *natural-sibling-of* relation: they are ordinary in the sense that they are to be interpreted as literal in all contexts. But a few 'extraordinary' relations, not commonly treated in kinship ontologies, appear on the CCA-01 form as well, including the *natural-older-uncle-of* relation. The Thai expression which appears on CCA-01 for this relation refers to a person who is an older biological brother of one of one's biological parents. These and other unusual relations we axiomatized in the separate module, cca01-ground.clif. The axioms therein are also to be interpreted literally – i.e. they represent universal ground truth – but are not useful in an environment in which such relations are not considered.

Three modules provide a bridge to and from SNOMED CT. Axioms linking from relations and universals to SNOMED CT concepts are in ancestry-sct.clif; those linking from SNOMED CT concepts to relations and universals are in sct-ancestry.clif. For example, ancestry-sct.clif contains an axiom to the effect that if x is natural parent of someone, then x is an individual of sct_Natural-parent_person; and sct-ancestry.clif contains an axiom to the effect that if x is an individual of sct_Natural-parent_person, then x is natural parent of someone. Whenever one of these modules is imported, the axioms in sct-declaration.clif also need to be imported. This module contains a collection of axioms, the first member of which states that if x is an individual of y, then x is a particular and y is a class, and the other members of which state about each term taken from SNOMED CT that it is a class. For example, one of the axioms in sct-declarations.clif states that sct_Natural-parent_person is a class.

The module unusual.clif contains 4 axioms pertaining to highly unusual situations, in which (1) a person z has a natural descendant which has as parent a natural ancestor of z; (2) a pair of close blood relatives are co-natural-parents of someone; (3) a pair of close blood relatives are spouses; and (4) some people are in a plural spousal situation. These unusual-case axioms can serve valuable data-checking purposes. If someone enters into an electronic medical record data that trigger one of these axioms (by, say, entering data to the effect that Fred and Sally are spouses while a contemporaneous spousal relation between Fred and Catherine is already on record), then a warning message can be devised recommending that the entered data be double-checked for accuracy. Given how unusual the situations in question are, the triggering data-entry will often have been erroneous.

For the sake of convenience, we also compiled a file, inspiration.clif, containing $T_{kinship}$ translated into CLIF but otherwise left untouched. This file is not to be considered part of our kinship ontology.

All axiom files as well as certain additional documentation is available via the following link: https://buffalo.box.com/s/pn9rv6m0i7wkcfow48f9270c4jh3kp6c

7. Discussion

Crucial to our ontology is the *has-natural-child* relation; the *natural-ancestor-of* relation is defined recursively partly in terms of it. Other blood relations are defined in our ontology partly in terms of one or the other of these two relations. By 'partly', we mean that additional information is needed to fully grasp the intended meaning. Although such information can be axiomatized as well, and should be done for other purposes extending kinship between individuals, it would not lead to useful reasoning in the context of CASCAP. By 'has-natural-child(Amy, Bob)', for example, we mean that among the gametes from which Bob originates is a gamete of Amy. By 'of Amy' in this context, we do not mean (say) *owned by Amy* or *controlled by Amy*, but rather *having its biological origin in Amy*. If Amy has sold one of her ova, O, to Clair, then O is in a loose sense a gamete *of Clair* but is not *of Clair* in the sense relevant here. The expression 'natural' is thus to be understood throughout our ontology as helping to designate blood relations. We use the term 'natural' in this context, as opposed to (say) 'biological' or 'blood', simply because 'natural' is also the term used in most SNOMED CT blood-relation concepts, and relevant modules in our ontology function as a bridge between SNOMED CT and BFO.

Many relations in our ontology, as in many kinship ontologies, are also partly defined in terms of sexes. In our axioms, the expression 'male-sex' picks out the sex had by males qua males, and 'female-sex' picks out the sex had by females qua females. We insist that 'male-sex' picks out the sex had by males qua males, not the sex had by males simpliciter (and similarly, mutatis mutandis, for 'female-sex' and females). Simultaneous hermaphrodites (e.g., great pond snails) are male and female at the same time. It follows that there is no such thing as the sex had by males simpliciter, for some males have multiple sexes. But even a simultaneous hermaphrodite has the male sex, and only the male sex, qua male. Hence our insistence on the "qua" restriction.

We also use the term 'spouse' to pick out a *marriage-partner*, and the expression 'has-spouse(x,y)' to mean that x has a marriage partnership with y. We take a rather minimal stand on the metaphysics of marriage – concerning who can enter into it, how many individuals can enter into a given marriage, whether marriage is 'purely legal' and so on. Our stand is not utterly neutral, though. For example, one axiom of ours entails that the *has-spouse* relation is irreflexive: you can't be married to yourself. We also assume that if x has spouse y, then there is a marriage bond that inheres in x and y and that exists at a time at which x and y exist. This assumption secures the intuitively correct verdict that spousehood is a *temporal* matter; people are not simply atemporally spouses of one another, even if natural-ancestry relations are simply atemporal. At present, our ontology says nothing about other non-blood relations beyond spousehood, such as domestic partnerships, close friendships, and so on; though it could be extended to comprehend such relations.

7.1. Time indexing

Because time-indexing plays such an extremely important role in BFO, a package of questions guiding our project was which elements of our ontology required time-indexing, which did not, and how the appropriate time-indexings would be best accomplished. The axioms in the ontologies discussed in Section 4, including $T_{kinship}$, treat (say) *being a person* as a matter of timelessly bearing some property, or as a matter of timelessly being, or being related in some way, to a member of some class. One feature of our ontology that makes it different from those discussed in Section 4 is thus that all such talk is replaced by talk of *individuals' being instances of universals at times*, as in, for example, the following axiom:

(A8) $\forall x \forall y(\text{natural-father-of}(x,y) \equiv$
$(\text{natural-parent-of}(x,y)$
$\& \exists q \exists t(\text{instance-of}(q,\text{male-sex},t)$
$\& \text{inheres-in}(q,x))))$

Relations between individuals included in our ontology were a somewhat trickier matter than universals, because for some of these relations time-indexing seems appropriate but for others it does not. For example, it is plausible that *natural-ancestor-of* is non–time-indexed: it seems atemporally true that, for example, Abraham Lincoln's maternal grandfather is among Lincoln's ancestors. By contrast, consider *has-spouse*. Some form of time-indexing seems appropriate for this relation, for people are married to one another for specific time periods, and a given person can be married to different people at different times. One way to accommodate an element of time-indexing concerning *has-spouse* is to time-index *has-spouse* itself; another is to attach time-indexing to something that one's ontology holds to be inextricable from spousal relations. As mentioned above, we took the latter approach, by maintaining that x has spouse y if and only if there is a marriage bond – a specialization of BFO2020's relational quality – that exists *when x and y do* and that inheres in x and y.

7.2. Bridging to SNOMED CT

The meaning of each SNOMED CT term is provided either through an individual concept or by at least one axiom expressed in the description logic EL++ [28]. Some of us have described elsewhere some of the potential benefits of using bridge axioms to attach the terminological richness of SNOMED CT to the ontological foundation supplied by BFO [23, 29]. The approach defended in [23, 29] is to let SNOMED CT's view and BFO's view happily co-exist, not in one *ontological* framework, but in one *logical* model-theoretic framework capable of exploiting what SNOMED CT offers *terminologically* and realism-based ontologies offer *ontologically*. We developed the modules sct-ancestry.clif, ancestry-sct.clif and sct-declarations.clif of our kinship ontology partly as a proof of concept of this general bridging strategy.

While devising our ontology, we encountered a number of challenges. For example, some of the relations at issue in ancestry.clif and cca01-ground.clif do not have precisely corresponding SNOMED CT concepts. For instance, one relation defined in cca01-ground.clif is the *natural-older-uncle-of* relation. Unsurprisingly, sct_Natural-older-uncle_person does not exist in the international version of SNOMED CT. Perhaps a bit more surprisingly, there is also no SNOMED CT concept precisely corresponding to the *natural-ancestor-of* relation. In these cases, bridging axioms cannot be proposed. When, however, a relation in our ontology lacked a precisely corresponding SNOMED CT concept, but there was a SNOMED CT concept that *nearly* precisely corresponded to the relation and a true bridge axiom connecting them could be imagined, we chose to include such an axiom. For example, because *natural-niece-of* and sct_Niece_person nearly precisely correspond and sct_Natural-niece_person does not exist, we included in our ontology the following bridging axiom:

(A9) $\forall$x($\exists$y(natural-niece-of(x,y)) $\rightarrow$ individual-of(x,sct_Niece_person))

A related issue concerned the question when bridging axioms could be formulated in both directions and when they could not be. In situations involving precisely corresponding kinship relations and SNOMED CT concepts, axioms in both directions were warranted; in situations not involving such precise correspondence, this was not the case. Hence the axioms (A10) and (A11), for example, both appear in our ontology, because *natural-sibling-of* and sct_Natural-sibling_person precisely correspond:

(A10) $\forall$x($\exists$y(natural-sibling-of(x,y))
$\rightarrow$ individual-of(x,sct_Natural-sibling_person))

(A11) $\forall$x(individual-of(x,sct_Natural-sibling_person)
$\rightarrow$ $\exists$y(natural-sibling-of(x,y)))

By contrast, the right-to-left analogue of (A9) does not appear in our ontology.

At present, our kinship ontology focuses primarily (though not exclusively) on blood-relations, and we have accordingly focused up to now on linking SNOMED CT blood-relative concepts to our ontology and proposing axioms defining relations corresponding to such concepts. However, there are a great many SNOMED CT blood-relative concepts, some of them highly specific, and we have not thus far tried to link every SNOMED CT blood-relative concept to our ontology or to propose axioms defining relations corresponding to every such concept. For example, sct_Identical-twin-brother_person falls under sct_Blood-relative_person, but we have not yet linked this concept to our ontology, we have not proposed an axiom defining the identical-twin-brother-of relation, and so on. Still, the work we have done thus far could be extended so as to cover all SNOMED CT blood-relative concepts, and indeed we hope in future work to do that.

7.3. Unusual kinship relations

A different stage of our project that required judgment calls on our part was the production of unusual.clif. Consider, for example, the following axiom in unusual.clif:

(A12) $\forall x \forall y \forall z$((has-natural-child(x,z) & has-natural-child(y,z) & (natural-parent-of(x,y) $\vee$ natural-sibling-of(x,y) $\vee$ natural-grandparent-of(x,y) $\vee$ natural-aunt-of(y,x) $\vee$ natural-uncle-of (y,x) $\vee$ natural-first-cousin-of(x,y)))
$\rightarrow$ occupy-unusual-ancestry-situation(x,y,z))

The intuitive idea behind (A12) is that if two people are co-natural-parents of a common person and are themselves close blood relatives, then they and their child occupy an unusual ancestry situation. The long disjunctive clause specifies a range of blood-relations that clearly qualify as *close*. But, of course, there are other close relations not covered by the clause (e.g., *natural-second-cousin-of*). Adding additional such relations to the disjunctive clause would allow for the generation of additional data-entry warnings, and so would probably flag some incorrect data-entries that would otherwise go undetected. However, expanding the disjunctive clause in certain imaginable ways might also generate enough warnings in cases of *correct* data-entry that doing so would not be all-things-considered prudent. For example, if one were to revise (A12) in such a way that co-natural-parenting natural seventh cousins count as occupying an atypical ancestry situation, then one would perhaps thereby revise (A12) in such a way that it yielded a counterproductively large number of false positives.

We suggest that (A12) be read as an *example* of a warning-case axiom relevant to the situations to which it pertains, not as *the best possible version* of such an axiom. This axiom and the others in unusual.clif could be revised to suit particular data-entry contexts, or even rejected altogether. For example, the axiom (A13) is in unusual.clif, we hope for obvious reasons:

(A13) $\forall x \forall y \forall z(\exists t \exists m1 \exists m2$(instance-of(m1,marriage-bond,t) & instance-of(m2,marriage-bond,t) & inheres-in(m1,x) & inheres-in(m1,y) & inheres-in(m2,x) & inheres-in(m2,z) & ~y=z) $\rightarrow$ occupy-unusual-spousal-situation(x,y,z))

However, if one lives in a polygamous society, then it might be a good idea for one not to adopt axiom (A13) at all. Though even this is not obviously right: if one lives in a society in which some but very few people practice polygamy, then (A13) might be worth adopting after all.

8. Conclusion

We have developed a novel kinship ontology in First Order Logic following the representational principles of BFO2020-FOL. The ontology comes in separate CLIF-modules each one of which can be imported based on specific needs, for example, mapping to and from SNOMED CT, or exploiting axioms which would not be literally true when phrased naively but are crafted in a way that allows the generation of alerts on possible data entry mistakes. The ontology can be used directly by CLIF-reasoners, or translated into much weaker versions of the axioms for OWL-DL reasoners. In future work, we intend to expand on this project by adding SNOMED CT kinship concepts that we have not yet wedded to our kinship ontology, either through definitions of corresponding relations or through production of relevant bridge axioms. Further expansion following the same bridging strategy might happen when other relevant kinship terminologies become prominently used.

Acknowledgements

This work was supported in part by Clinical and Translational Science Award UL1 TR001412 from the National Institutes of Health and a T15 grant awarded by the National Library of Medicine. We would like to thank Dr. Kavin Thinkhamrop for providing valuable information about CASCAP.

References

[1] Ochs C, Perl Y, Geller J, Arabandi S, Tudorache T, Musen MA. An empirical analysis of ontology reuse in BioPortal. J Biomed Inform. 2017;71:165-77. doi: 10.1016/j.jbi.2017.05.021. PMID: 28583809; PMC5557647.

[2] Jackson RC, Balhoff JP, Douglass E, Harris NL, Mungall CJ, Overton JA. ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics. 2019;20(1):407. Epub 20190729. doi: 10.1186/s12859-019-3002-3. PMID: 31357927; PMC6664714.

[3] Rector A, Schulz S, Rodrigues JM, Chute CG, Solbrig H. On beyond Gruber: "Ontologies" in today's biomedical information systems and the limits of OWL. J Biomed Inform. 2019;100S:100002. Epub 20190309. doi: 10.1016/j.yjbinx.2019.100002. PMID: 34384571.

[4] Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The Human Disease Ontology 2022 update. Nucleic Acids Res. 2022;50(D1):D1255-D61. doi: 10.1093/nar/gkab1063. PMID: 34755882; PMC8728220.

[5] Ren Y, Parvizi A, Mellish C, Pan JZ, van Deemter K, Stevens R, editors. Towards Competency Question-Driven Ontology Authoring. The Semantic Web: Trends and Challenges; 2014 2014//; Cham: Springer International Publishing.

[6] Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology. Cambridge, Massachusetts: Massachusetts Institute of Technology; 2015. xxiv, 220 pages.

[7] Smith B, Ceusters W. Ontological realism: A methodology for coordinated evolution of scientific ontologies. Appl Ontol. 2010;5(3-4):139-88. doi: 10.3233/AO-2010-0079. PMID: 21637730; PMC3104413.

[8] International Standards Organisation. ISO/IEC 21838-2:2021 Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO). 2021.

[9] Alsaleh M, Leftley Z, Barbera TA, Sithithaworn P, Khuntikeo N, Loilome W, et al. Cholangiocarcinoma: a guide for the nonspecialist. Int J Gen Med. 2019;12:13-23. Epub 2018/12/28. doi: 10.2147/ijgm.S186854. PMID: 30588065; PMC6304240.

[10] Sripa B, Pairojkul C. Cholangiocarcinoma: lessons from Thailand. Curr Opin Gastroenterol. 2008;24(3):349-56. doi: 10.1097/MOG.0b013e3282fbf9b3. PMID: 18408464.

[11] Sripa B, Kaewkes S, Sithithaworn P, Mairiang E, Laha T, Smout M, et al. Liver fluke induces cholangiocarcinoma. PLoS Med. 2007;4(7):e201. Epub 2007/07/12. doi: 10.1371/journal.pmed.0040201. PMID: 17622191; PMC1913093.

[12] Khuntikeo N, Loilome W, Thinkhamrop B, Chamadol N, Yongvanit P. A Comprehensive Public Health Conceptual Framework and Strategy to Effectively Combat Cholangiocarcinoma in Thailand. PLoS Negl Trop Dis. 2016;10(1):e0004293. Epub 2016/01/23. doi: 10.1371/journal.pntd.0004293. PMID: 26797527; PMC4721916.

[13] Khuntikeo N, Chamadol N, Yongvanit P, Loilome W, Namwat N, Sithithaworn P, et al. Cohort profile: cholangiocarcinoma screening and care program (CASCAP). BMC Cancer. 2015;15:459. Epub 2015/06/10. doi: 10.1186/s12885-015-1475-7. PMID: 26054405; PMC4459438.

[14] Chamadol N, Pairojkul C, Khuntikeo N, Laopaiboon V, Loilome W, Sithithaworn P, et al. Histological confirmation of periductal fibrosis from ultrasound diagnosis in cholangiocarcinoma patients. Journal of Hepato-Biliary-Pancreatic Sciences. 2014;21(5):316-22. doi: 10.1002/jhbp.64.

[15] Cholangiocarcinoma Foundation of Thailand. Isan Cohort Khon Kaen University, Thailand: CASCAP: Cholangiocarcinoma and Care Program; 2016 [July 9, 2020]. Available from: https://cloud.cascap.in.th/.

[16] Kirstein MM, Vogel A. Epidemiology and Risk Factors of Cholangiocarcinoma. Visc Med. 2016;32(6):395-400. Epub 20161201. doi: 10.1159/000453013. PMID: 28229073; PMC5290446.

[17] Liu ZY, Zhou YM, Shi LH, Yin ZF. Risk factors of intrahepatic cholangiocarcinoma in patients with hepatolithiasis: a case-control study. Hepatobiliary Pancreat Dis Int. 2011;10(6):626-31. doi: 10.1016/s1499-3872(11)60106-9. PMID: 22146627.

[18] Chui C, Grüninger M, Wong J. An Ontology for Formal Models of Kinship. Fr Art Int. Frontiers in Artificial Intelligence and Applications. 3302020. p. 92-106.

[19] Stevens R, Sattler U, Stevens M. A Family History Knowledge Base in OWL 2. In: Bail S, Glimm B, E., Matentzoglu N, Parsia B, Steigmiller A, et al., editors.: RWTH Aachen University; 2014. p. 71-6.

[20] Rehm HL, Page AJH, Smith L, Adams JB, Alterovitz G, Babb LJ, et al. GA4GH: International policies and standards for data sharing across genomic research and healthcare. Cell Genom. 2021;1(2). doi: 10.1016/j.xgen.2021.100029. PMID: 35072136; PMC8774288.

[21] Orion Buske, Jim Balhoff, Michael Franklin, Chris Mungall. KIN - Family History Terminology 2022. Available from: https://github.com/GA4GH-Pedigree-Standard/family_history_terminology/blob/main/src/main/resources/kin.owl.

[22] Longo C, Gangemi A, Cantone D, editors. Representing Kinship Relations on the Semantic Web. OWL: Experiences and Directions; 2013.

[23] Anuwat Pengput, Werner Ceusters. Setting the Scene to Link SNOMED CT to Realism-Based Ontologies. 19th world congress on medical and health informatics (MEDINFO 2023); Jul 11, 2023; Sidney, Australia, 2023.

[24] Smith B. Against Fantology. In: Reicher ME, Marek JC, editors. Experience and Analysis. Wien2005. p. 153-70.

[25] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005;6(5):R46. Epub 2005/05/17. doi: 10.1186/gb-2005-6-5-r46. PMID: 15892874; PMC1175958.

[26] Anuwat Pengput, Alex Diehl. Development of Axiomatization for the Cholangiocarcinoma Ontology Using Common Logic Interchange Format. International Conference on Biomedical Informatics (ICBO); Ann Arbor, MI, USA2022.

[27] Kowalski R. Predicate Logic as a Programming Language. Proceedings of IFIP-74. Amsterdam, The Netherlands: North-Holland; 1974. p. 569-74.

[28] Rodrigues JM, Schulz S, Mizen B, Trombert B, Rector A. Scrutinizing SNOMED CT's Ability to Reconcile Clinical Language Ambiguities with an Ontology Representation. Stud Health Technol Inform. 2018;247:910-4. Epub 2018/04/22. PMID: 29678093.

[29] Ceusters W, Pengput A. Axiomatizing SNOMED CT Disorders: Should there be Room for Interpretation? Formal Ontology in Information Systems (FOIS) 2023; Sherbrooke, Quebec, Canada: (in press); 2023.

Towards principles of ontology-based annotation of clinical narratives

Stefan Schulz[1,2,*], Warren Del-Pinto[3], Lifeng Han[3], Markus Kreuzthaler[1], Sareh Aghaei[1] and Goran Nenadic[3]

[1] *Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria*
[2] *Averbis GmbH, Freiburg, Germany*
[3] *Department of Computer Science, University of Manchester, UK*

Abstract

Despite the increasing availability of ontology-based semantic resources for biomedical content representation, large amounts of clinical data are in narrative form only. Therefore, many clinical information management tasks require to unlock this information using natural language processing (NLP). Clinical corpora annotated by humans are crucial resources. On the one hand, they are needed to train and domain-fine-tune language models with the goal to transform information from unstructured free text into an interoperable form. On the other hand, manually annotated corpora are indispensable for assessing the results of information extraction using NLP. Annotation quality is crucial. Therefore, detailed annotation guidelines are needed to define the form that extracted information should take, to prevent human annotators from making erratic annotation decisions and to guarantee a good inter-annotator agreement. Our hypothesis is that, to this end, human annotations (and subsequently machine annotations learned from human annotations) should (i) be based on ontological principles, and (ii) be consistent with existing clinical documentation standards. With the experience of several annotation projects, we highlight the need for sophisticated guidelines. We formulate a set of abstract principles on which such guidelines should be based, followed by examples of how to keep them, on the one hand, user-friendly and consistent, and on the other hand compatible with the international semantic standards SNOMED CT and FHIR, including their areas of overlap. We sketch the representation of the resulting representations in a knowledge graph as a state-of-the-art semantic representation paradigm, which can be enriched by additional content on A-Box and T-Box levels and on which symbolic and neural reasoning tasks can be applied.

Keywords

Formal Ontologies, Clinical Information Models, Natural Language Processing, Text Annotation Guidelines, Electronic Health Records

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*] Corresponding author

stefan.schulz@medunigraz.at (S. Schulz); warren.del-pinto@manchester.ac.uk (W. Del-Pinto); lifeng.han@manchester.ac.uk (L. Han); markus.kreuzthaler@medunigraz.at (M. Kreuzthaler); sareh.aghaei-dinani@medunigraz.at (S. Aghaei); gnenadic@manchester.ac.uk (G. Nenadic)
0000-0001-7222-3287 (S. Schulz); 0000-0003-3307-9432 (W. Del-Pinto); 0000-0002-3221-2185 (L. Han); 0000-0001-9824-9004 (M. Kreuzthaler); 0000-0002-0511-095X (S. Aghaei); 0000-0003-0795-5363 (G. Nenadic)

CEUR Workshop Proceedings (CEUR-WS.org)

1. Introduction

A seamless and effective flow of clinical information is vital for high-quality healthcare and health management. Thus, information must be stored in a way that supports effective communication, search, and analysis. However, most content of electronic health records (EHRs) consists of unstructured text in documents and free-text database fields. As a result, crucial data that require key insights into populations and individuals remain inaccessible without additional processing.

In contrast, the good news regarding EHR interoperability is the increasing support by elaborated semantic standards, *viz.* terminologies, ontologies (e.g., SNOMED CT, LOINC) [1] and information models (e.g., FHIR [2]), which enjoy increasing international adoption. Interoperable clinical data representations would ideally bridge syntactically different but semantically equivalent expressions in different degrees of structure, from data collected in standardised forms to clinical narratives, across disciplines, jurisdictions, and natural languages.

To achieve this goal for narrative data, text passages must be linked to identifiers from a controlled vocabulary, a process referred to as semantic annotation or tagging. Such vocabularies are typically rooted in semantic resources such as those mentioned above. An additional step is the assertion of links between tags, known as relation annotation. To do this manually is resource-intensive and difficult in practice; human annotators need to be trained and monitored, and diverging annotations must be reconciled by subsequent adjudication.

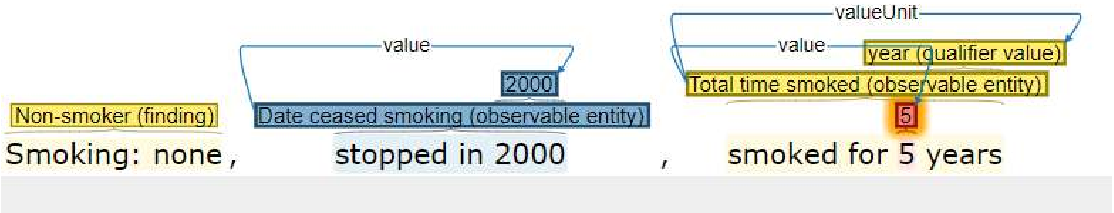
Nevertheless, annotated text corpora are indispensable as a "fuel" for training, domain-fine-tuning, and evaluation of natural language processing (NLP) models. Fig. 1 shows an example of a manual annotation of a clinical text passage.

Annotation guidelines play a vital role in this process. They target consistency and uniformity by providing clear instructions, reducing variability, and ensuring that annotations are standardised across annotators and projects. The inter-annotator agreement serves as a measure of the effectiveness and quality of guided annotation processes. Annotation guidelines provide instructions on, and examples of, handling ambiguous cases, addressing recurring challenges, and resolving annotation discrepancies.

An ideal annotation should produce, with the same input text, the same target representation created by different trained annotators. The same should be achieved based on different paraphrases of the same input or with its translation into a different language. Even where annotations differ, semantic equivalence should be stated based on logical reasoning. Although these desiderata will probably never be completely fulfilled, they justify more effort spent on principled annotation guidelines.

Several guidelines for annotating clinical narratives have been proposed. Examples include the Clinical E-Science Framework (CLEF) [3, 4] which focuses on clinical information extraction tasks, annotation guidelines for the identification of personal health information such as patient names [5] and a range of guidelines that are very specific in terms of language and task, e.g. lung diseases in Japanese [6], family history in Norwegian [7], diseases in Spanish [8], disorders, findings, drugs and body structures in Swedish and Chinese [9, 10]. Guidelines may also specify how to annotate temporal references [11, 12], negation statements [13], and other contextual markers.

Although the recognition of diagnostic statements has been a preferred goal in clinical annotation tasks, the need for a broader coverage of annotation guidelines has been recognised


valueUnit
value
value
year (qualifier value)
2000
Total time smoked (observable entity)
Non-smoker (finding)
Date ceased smoking (observable entity)
5

Smoking: none , stopped in 2000 , smoked for 5 years

Figure 1: Semantic annotation example of a passage from a discharge summary

by Luo et al. [14], who extended the scope of annotation to "medical problems, treatments and tests". These examples show that previous works on annotating clinical text have often produced guidelines that were specific in scope, constrained to a particular semantic type or designed for a very use-case-specific data set.

Although the argument that specific use cases necessitate specific annotation strategies is valid, we postulate a need for more general principles that provide a systematic approach to clinical text processing. Such an approach should be based on ontological principles [15, 16] and international semantic standards [17]: a prerequisite to ensuring that data from different sources can be meaningfully compared. Once such principles have been formulated, task-specific guidelines could be instantiated without unnecessary repetition of previous efforts.

In this work, we propose such general annotation principles. Regarding the annotation vocabulary, we focus on SNOMED CT [18], an ontology-based clinical terminology system, and FHIR [2], a set of information templates for recurring clinical documentation tasks. Both are considered leading international semantic standards for healthcare data. We separate ontological aspects – the description of classes of biomedical entities and their properties, which is the domain of SNOMED CT – from contextual information about individuals, which are represented as instances of FHIR resources. This addresses A. Rector's claim to distinguish "models of meaning" (ontologies) and "models of use" (information models) [19]. Applied to clinical text annotation, this means that FHIR resources provide semantically explicit templates to capture instance-level information on epistemic, temporal, and provenance aspects belonging to a given patient and his/her documentation context. In contrast, SNOMED CT codes represent types of clinical entities[1], whose instances are referred to by an appropriate field in the FHIR template. It is well known that SNOMED CT also supports epistemic and temporal aspects in its *Situation in specific context* hierarchy, whereas FHIR resources include references to HL7 value sets, mostly for roles and qualities, which compete with SNOMED CT *Qualifier value* concepts. This overlap constitutes a complicating factor of clinical information management in general and clinical text annotations in particular. Efforts to principally handle this overlap[2] have remained unfinished.

[1] AKA concepts or T-box entities, modelled as OWL classes.

[2] TermInfo Project. (This and the following footnotes: click to navigate to the respective project website)

2. Methods

We employed a qualitative and heuristic approach to develop annotation guidelines for standards-aware representation of clinical documents. The goal was to bootstrap annotation rules from sample narratives, inspect the results, discuss disagreements between annotators, identify recurring patterns and structures, and consolidate a final guideline after a series of iterations. The authors have gained experience in numerous annotation activities. These include the use of SNOMED CT for the annotation of clinical summaries in Brazilian Portuguese [20] and for clinical text snippets in five European languages [21], and the application of a set of manually designed annotation rules to TAC2017[3] data to supervise the extraction of adverse drug reactions and related entities. SNOMED CT has also been used to provide the semantics for normalising clinical event mentions in diagnostic statements extracted from hospital data in the UK.

The problems of representing context using SNOMED CT and FHIR [22] led the authors to pilot annotation tasks using the German corpus GRASCCO[4] and a clinical corpus derived from UK hospital data by using FHIR, including HL7 value sets, as a framework in which SNOMED CT is used for the ontological information proper. Soon, it became necessary to formulate rules to manage the overlap between both systems. In addition, it became clear that the handling of the complex structure of FHIR was difficult for the human annotators.

Guideline-based text annotations are also in the centre of the ongoing projects AIDAVA[5] (Horizon Europe), the German annotation initiative GemTeX[6], as well as the UK projects JIGSAW[7] and HIPS[8]. The latter projects required the formulation of annotation guidelines to create diagnosis representations in SNOMED CT, using the *Situation in specific context* hierarchy, which proposes precoordinated content and post-coordination patterns for factuality (e.g. “suspected asthma”), temporality (e.g. “history of major depression”) and family history (“mother died from breast cancer”).

While FHIR is compatible with multiple terminologies, the use of SNOMED CT was motivated in part by its widespread adoption, e.g., by the requirement that all systems of the UK National Health Service (NHS) must use SNOMED CT as a core terminology[9]. Additionally, given the required annotation tasks, the broad coverage of SNOMED CT hierarchies, which include clinical findings (disorders), pharmaceutical products, procedures, and others, was deemed beneficial.

The continuous crafting and revision of annotation guidelines, supported by the outcomes of annotator training sessions, resulted in the creation of a set of annotation principles as formulated in the next section, followed by a case study that serves as an instantiation of these principles, centring on the rooting of annotations into the standards SNOMED CT and FHIR.

[3] Adverse Drug Reaction Extraction from Drug Labels

[4] Graz Synthetic Clinical Text Corpus

[5] AI-powered Data Curation & Publishing Virtual Assistant

[6] German Medical Text Corpus

[7] Assembling the Data Jigsaw

[8] Healthcare Impact Partnership: Integrating hospital outpatient letters into the healthcare data space

[9] NHS Digital, SCCI0034

3. Results

3.1. Proposed Annotation Principles

Different use cases require different annotation guidelines. For example, secondary uses of clinical data such as epidemiological research have different requirements than data representation for providing direct patient care or data support for billing purposes. Therefore, we first propose a set of general principles from which more specific annotation guidelines for given applications can be derived. This first set of annotation principles is independent of the annotation vocabularies used, i.e. the ontologies and information models (or subsets thereof) used.

- Annotation is limited to the *semantic* aspects of narratives. It comprises the assignment of codes for unary predicates (types, but also individuals) from a domain ontology, together with literals such as decimal numbers to spans of text. These annotations are additionally connected by binary predicates (relations). Both unary and binary predicates are rooted in a semantic reference such as a domain ontology or a clinical information model.
- The endpoint is a canonical form of representing clinical narrative information as a primary knowledge graph (KG), with subject-predicate-object triples describing individual patients and related clinical entities. This primary KG should then be transferable, by applying supporting rules and resources, into a knowledge graph that follows the structures of the underlying standards and is committed to Applied Ontology principles. Such a KG makes a clear T-Box/A-Box distinction, i.e. between nodes that point to entity types as given by an ontology and those that point to individual entities that instantiate the types provided by the annotation. The difference is relevant, because, e.g., “asthma” should point to an individual entity (particular) in an affirmative context (“The patient has asthma”), but to an entity type (concept) in a negative context (“no signs of asthma”). Another ontological distinction is the one between information entities [23] (instances of FHIR resources) and the clinical entities they are about.
- This complexity needs to be hidden from annotators. Predicate types should be restricted to the necessary. E.g., the relation between “tumour” and “right” is ontologically very different from the one between “tumour” and “suspected”. Nevertheless, a literal annotation (by using predicates such as “laterality” and “verificationStatus” in the same way) would suffice for the annotators, provided that annotation post-processing is sufficiently specified by transformation rules.
- Scope and granularity are determined by the underlying annotation vocabulary. For restricted annotation tasks, subsets of the maximum vocabulary are provided to the annotators. For example, a targeted cancer annotation task might not require fine-grained reference to unrelated diseases like heart disorders or injuries. This could motivate the pruning of sub-hierarchies, e.g., underneath *Coronary disease* or *Fracture of bone*.
- Annotations are descriptive and not interpretative: annotators annotate only what they read, without complicating their task by seeking individual interpretations. For example, the annotations “fever” after “hip replacement” are only linked with a predicate for causality if there is a causality statement in the text. Two exceptions to this rule are highlighted: (i) word sense disambiguation, as long as meaning can be derived from the context; and

(ii) co-reference, as long as the antecedent to which an anaphoric reference points back is identifiable.

- The granularity of annotation spans is not given by a named entity recognition step prior to annotation, which would yield entity types, such as *Disorder* and *Body part* as annotations of the spans "fracture of skull" or "left ventricle". Instead, the spans are determined by the underlying annotation vocabulary. The principle of longest match is followed and pre-coordinated expressions are used as preferred if they correspond to a contiguous span. Otherwise, e.g. in "the skull exhibited the sign of a fracture", shorter text spans ("skull", "fracture") are annotated and linked afterwards. If the representation of the meaning of a span requires more than one code, a conjunction (logical AND) is preferred over a construction that additionally requires binary predicates.
- The annotation vocabulary is distinguished between *core* content and *supporting* content. The former one characterises the central focus of the given annotation task, as defined by the intended use case, while the latter one provides additional, mostly refining information to the core. For example, for the identification of diagnosis statements, core content would be given by the concepts of the Clinical Finding hierarchy of SNOMED CT, such as diseases. Supporting concepts would be those under *Body Structure*, which specify a particular *Clinical finding* concept, as well as those that capture factuality, such as *Probably present*.

The following annotation principles explicitly refer to the use of SNOMED CT and HL7 FHIR as annotation vocabularies.

- "Core" hierarchies as introduced above are *Clinical finding*, *Event*, *Observable entity*, *Pharmaceutical / biologic product*, *Procedure*, *Specimen* in SNOMED CT. They have a high proportion of fully defined concepts, expressed by OWL 'equivalentTo' axioms, in which concepts from "non-core" hierarchies such as *Substance*, *Organism*, *Body structure* and *Physical object* are referred to by existentially quantified links. Ambiguous annotations are addressed by preference rules, e.g. to prefer C_1 over C_2, if C_1 belongs to a core hierarchy. For example, SNOMED CT offers different codes for "Sarcoma", *viz. Sarcoma (disorder)* and for *Sarcoma (morphological abnormality)*. The former is preferred for annotation of clinical texts because it is fully defined and axiomatically implies the latter.
- The hierarchy *Situation in specific context* – although it would formally correspond to a core hierarchy – is not used for annotation because FHIR has been shown to be more granular, actively maintained, and frequently used to represent the context of clinical statements.
- A set of binary predicates with their own namespace "anno:" was introduced for close-to-user relation annotation. These binary predicates were grounded in (i) SNOMED CT object properties or chains thereof, (ii) relational chains of FHIR elements, or (iii) both. For example, the predicate **site** between a SNOMED CT clinical finding and a body structure, is mapped to the linkage concept (relation) **'finding site'** as well as the concatenation of the inverse of the FHIR element **Condition.code** with **Condition.body**, cf. Table 2.
- SNOMED CT mappings to HL7 value sets are proposed, e.g. *hl7:Recurrence* to *sct:Recurrent* or *hl7:Confirmed* to *'sct:Confirmed present'* (see Table 2).

Guidelines designed for a specific annotation task can be viewed as an instantiation of the general principles outlined above. It is then necessary to define sets of permitted codes for both the core and supporting concepts to be drawn from, for example, SNOMED CT reference sets. To summarise, the steps in utilising such a guideline to annotate clinical text are as follows:

- Identify a core concept mentioned explicitly in the document.
- Assign to the identified phrase a suitable concept from the set of possible core concepts.
- If present, identify phrases corresponding to supporting concepts that refine the core concept.
- Assign to each identified supporting concept a suitable SNOMED CT concept, drawn from the set of possible supporting concepts.
- For each of the supporting concepts, link them directly to the core concept and identify the type of relation.

Given a clinical document, the above steps can be repeated until all of the relevant clinical mentions have been captured. The final annotation shall be a semantic representation of the explicit meaning of the original clinical text.

3.2. Examples

Here, we provide examples of applying these general principles to produce task-specific templates for annotating clinical text.

3.2.1. Example from JIGSAW/HIPS

The JIGSAW and HIPS projects include the task to extract diagnosis information from both semi-structured lists and free-text narratives in outpatient letters for secondary use purposes such as specifying patient cohorts for epidemiological studies. The annotation task was decided to be uniform across both projects to ensure consistent capture of relevant information. The combination of SNOMED CT with FHIR naturally resulted from the NHS use of SNOMED CT as its preferred terminology, and the need to represent and communicate instance-level information about patients. Following the principles outlined above, core and supporting concepts were identified. As a source for core concepts all concepts in the SNOMED CT *Clinical Finding* hierarchy were taken. The supporting concepts were primarily taken by the elements of the FHIR Condition resource[10], with additional relations from the SNOMED CT concept model including *Associated morphology* and *Causative agent*. In order to keep the annotation task consistent and simple, the initial assignment of codes to diagnostic statements in the text has been entirely done using concepts from the SNOMED CT *Clinical finding* hierarchy. The linking of supporting concepts to their corresponding core concepts follows the strategy outlined in the previous section: a set of binary predicates were specified for relation annotation, to avoid the need for annotators to be familiar with all SNOMED CT "linkage concepts"[11] and FHIR elements. In this context, given the narrative phrase "osteoarthritis of the spine", the FHIR Condition resource specifies that

[10] FHIR Condition

[11] Corresponding to OWL object and datatype properties.

a *Clinical finding* code should be provided for the condition being diagnosed and, if necessary, a body part can also be provided in the form of a SNOMED CT *Body structure* code. Therefore, *Osteoarthritis* is a core concept that is refined by the SNOMED CT *Clinical finding* concept *Osteoarthritis*. Meanwhile, "spine" acts as supporting information. It is annotated using the *Body Structure* concept *Joint structure of spine*. As a result, it is necessary to perform post-processing both to map the provided SNOMED CT codes to appropriate FHIR values and to map the relations to either SNOMED CT object properties or FHIR elements where applicable.

For FHIR elements with value sets that are specified as SNOMED CT codes by default, such as **Condition.bodySite**, no mapping was required. For those that specify an alternative FHIR value set, such as **Condition.verificationStatus** and **Condition.clinicalStatus**, mappings were specified between appropriate SNOMED CT concepts and the FHIR values. These mappings were based upon discussion with clinicians regarding which information is needed for their use case, such as the need to specify uncertainty of diagnoses via SNOMED CT concepts such as *Probably present*, and how these should be interpreted in FHIR, for example *Provisional*.

3.2.2. Example from AIDAVA

Regarding the annotation of composed expressions, AIDAVA puts more emphasis on the use of pre-coordinated SNOMED CT content. For the contiguous span "osteoarthritis of the spine" preference would be given to the single concept *Spondylosis*, whereas a passage such as "osteoarthritis of knees, hips, spine" would require a post-coordinated expression such as exemplified in 3.2.1. The important is here that due to the formal axioms in SNOMED CT, equivalence can be stated between pre-coordinated content and post-coordinated expressions.

AIDAVA also required mappings between SNOMED CT and FHIR as shown in Table 1. In order to support, alternatively, queries on SNOMED CT and FHIR, more attention was paid to the interoperability of predications, i.e. the use of binary predicates and their anchoring in both standards. In general, predications are straightforward at a text level but complex at an ontology-based representation level. The phenomenon that representations may compete, e.g. SNOMED CT only vs. SNOMED CT in FHIR contexts had to be addressed, cf. Table 2. Figure 2 demonstrates the generation of an ontology-based knowledge graph, as intended by AIDAVA, consisting of SNOMED CT concepts, instances thereof, and literals like dates and identifiers, emerging from guideline-driven annotations of clinical narratives. It shows how, according to the annotation predicates chosen, different FHIR resources, *viz. Condition* and *FamilyMemberHistory* are instantiated. It also shows how the choice of the predicate **anno:verificationStatus** related to *sct:Suspected* leads to a reference to the concept '*sct:Neoplasm of breast*', whereas the annotation without modifier (which assumes that the diagnosis is confirmed) the same FHIR relation points to an *individual referent* that instantiates '*sct:Neoplasm of breast*'. Other details are only hinted at, such as the class *Information content entity* from the Information Artifact Ontology[12], as well as the inferred relation '**occurs in**' from the Basic Formal Ontology (BFO)[24].

[12] Information Artifact Ontology (IAO)

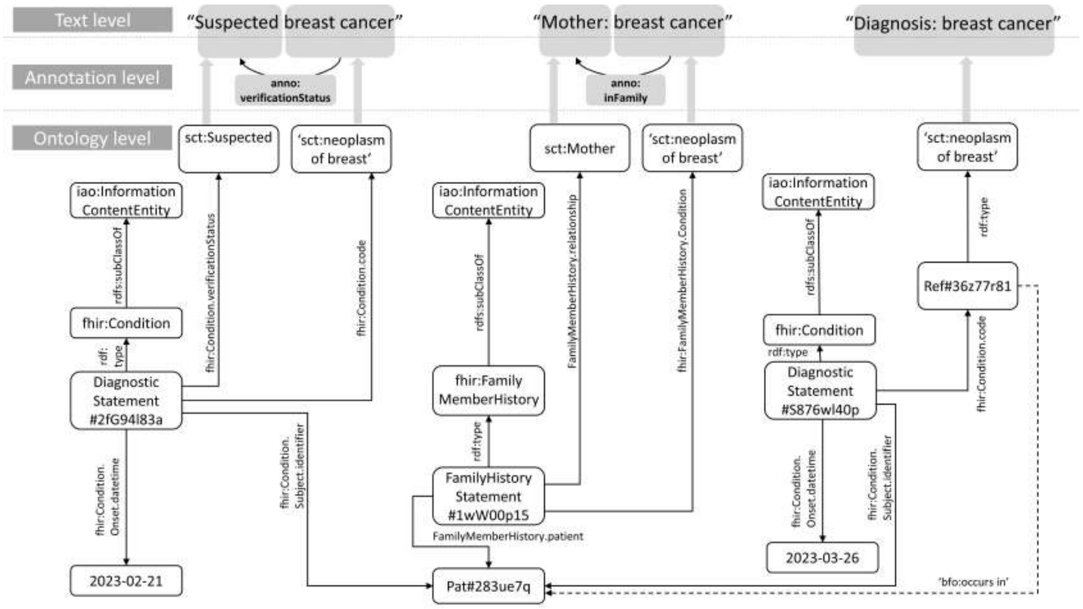

Text level
"Suspected breast cancer"
"Mother: breast cancer"
"Diagnosis: breast cancer"
Annotation level
anno: verificationStatus
anno: inFamily
Ontology level
sct:Suspected
'sct:neoplasm of breast'
sct:Mother
'sct:neoplasm of breast'
'sct:neoplasm of breast'
iao:Information ContentEntity
rdfs:subClassOf
fhir:Condition
rdf: type
Diagnostic Statement #2fG94l83a
fhir:Condition.verificationStatus
fhir:Condition.code
fhir:Condition. Onset.datetime
fhir:Condition. Subject.identifier
2023-02-21
iao:Information ContentEntity
rdfs:subClassOf
fhir:Family MemberHistory
rdf:type
FamilyHistory Statement #1wW00p15
FamilyMemberHistory.relationship
fhir:FamilyMemberHistory.Condition
FamilyMemberHistory.patient
Pat#283ue7q
iao:Information ContentEntity
rdfs:subClassOf
fhir:Condition
rdf:type
Diagnostic Statement #S876wl40p
fhir:Condition. Onset.datetime
fhir:Condition. Subject.identifier
2023-03-26
rdf:type
Ref#36z77r81
fhir:Condition.code
'bfo:occurs in'

Figure 2: From text to ontology-based knowledge graphs. The text level shows three text snippets (quoted) belonging to clinical texts about one patient. The human annotations (grey background) use SNOMED CT and "anno:" as annotation vocabularies. The lower, most complex level shows the instantiations of FHIR resources with references to SNOMED concepts, instances thereof, and literals. Nodes with random IDs represent individuals, linked to the concept whose code appears in the annotation. An example of an inferred predication is shown by a dashed arrow.

Table 1
Examples for mappings between HL7 FHIR values and SNOMED CT concepts

HL7 FHIR	SNOMED CT (from the *Qualifier value* hierarchy)
Unconfirmed	*Suspected* OR *Probably not present*
Provisional	*Probably present* OR *Suspected*
Confirmed	*Confirmed present*
Refuted	*Known absent*

4. Discussion

Past clinical annotation projects were often based on UMLS CUIs, as freely accessible but semantically often shallow concept identifiers, or on in-house annotation languages which restrict their openness such as in [10]. In other cases, annotations were limited to high-level entity types, such as *Disorder* and *Body part*, with a focus on relations. Many different ontologies were used, among which the Human Phenotype Ontology (HPO) or the Disease ontology (DO) should be emphasised.

The reason why we focus on SNOMED CT is its growing international acceptance as a standard for all health record content, its scope and granularity, and particularly its logical under-fitting, which facilitates the bridging between pre-coordinated and post-coordinated expressions.

However, our approach also have some reservations. It has been argued that SNOMED CT is little used in routine, particularly in continental Europe, that current licenses exclude important jurisdictions, and that translations are still missing. We reply that the status quo in clinical terminologies, with national ICD versions, national procedure classifications and drug catalogues, does not offer a convincing interoperability perspective without SNOMED CT. Furthermore, SNOMED CT is fully available to the research community, so the fact that its productive use requires a licence, should not be an obstacle.

One specific limitation is the still unresolved management of the overlap between SNOMED CT, FHIR, and related value sets. A further continuation of the Terminfo work in the light of FHIR would be desirable. Another limitation is that numerous SNOMED CT concepts lack formal and textual definitions, and pose challenges to annotators, particularly with texts in languages for which no official translation exists.

We are convinced that in times when large language models are skyrocketing, and under the hypothesis that machine understanding of clinical language is a realistic goal, semantic standards do not become obsolete. On the contrary, large language model technology has to be leveraged to generate canonical, standardised representations. Such representations as a gold standard for clinical content representation need to be elaborated and refined. We understand the proposed annotation principles as a step in this direction.

5. Conclusion and Future Work

Clinical annotation guidelines are crucial for structured data preparation, content representation for human reading and searching, enhancing supervised learning and for benchmarking language models via gold-standard data sets. Clinical annotation tasks become even more sophisticated due to the diversity of the text and broad coverage of knowledge across multiple dimensions, such as diseases, signs, symptoms, findings, procedures, as well as temporality, factuality, and other

Table 2
Example binary predicates for relation annotations with their translation into SNOMED CT and FHIR syntax. “INV” = inverse, “||” = concatenation. The relation paths marked with [a] refer to the concatenation of SNOMED CT relations, those marked with [b] to roughly equivalent FHIR elements

anno:	Domain	Target path	Range
site	‘sct:Clinical finding’	[a]‘sct:Finding site’ [b] INV(fhir:Condition.code) || fhir:Condition.body	‘sct:Body structure’
site	‘sct:Procedure’	[a]‘sct:Procedure - Direct’ [b] INV(fhir:Procedure.code) || fhir:Procedure.body	‘sct:Body structure’
inFamily	‘sct:Clinical finding’	[b] INV(fhir:FamilyMemberHistory.condition) || fhir:FamilyMemberHistory.relationship [a] INV(‘sct:Associated finding’) || ‘sct:Subject relationship context’	‘sct:Person’
verification status	‘sct:Clinical finding’	[b] INV(fhir:Condition.code) || fhir:Condition.verificationStatus [a] INV(‘sct:Associated finding’) || ‘sct:Finding context’	Qualifier value’ (cf. Tab. 1)

contexts. Existing annotation guidelines have been mostly motivated by shared task organisers to solve NLP challenges such as entity recognition and relation extraction, which tended to lead to shallow policies. With the purpose to address clinical annotations from an ontology-driven methodology and to take advantage of the rich content of SNOMED CT and FHIR, we proposed a set of annotation principles and designed the mapping of annotations into SNOMED CT and FHIR via entity and relation linking, as well as expression normalisation. The full design and implementation of our annotation principles cover a broad and in-depth semantic representation of the original clinical text, for which we recommend a representation as knowledge graphs, which can then be enriched by additional content on A-Box and T-Box level and on which both symbolic and neural reasoning tasks can be applied. For practical applications, we do suggest the users carry out the core annotation first and choose the level of depths of annotations according to their needs.

We understand that clinical free text annotation is a huge and challenging task, and the goal of achieving our full guideline principles might take a long journey. But there is indeed such a need to unify the clinical annotation tasks so that it can facilitate clinical research across different sectors, and languages, as well as for current large NLP model training.

In future work, we will prepare more annotation examples with a full set of instructions. We also plan to carry out the evaluation of our principles from the NLP perspective, in addition to case studies for measuring the inter-rater agreement [25] levels via trained workers.

6. Acknowledgments

The work was partially supported by Grant 101057062 “AIDAVA” (funder: the European Commission, HORIZON-HLTH-2021), Grant “Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease” (The project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation. Visit www.nuffieldfoundation.org) and Grant EP/V047949/1 “Integrating hospital outpatient letters into the healthcare data space” (funder: UKRI/EPSRC).

References

[1] O. Bodenreider, R. Cornet, D. J. Vreeman, Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm, Yearb Med Inform 27 (2018) 129–139.

[2] D. Bender, K. Sartipi, HL7 FHIR: An agile and RESTful approach to healthcare information exchange, in: Proc 26th IEEE Symp on Computer-based Med Syst, 2013, pp. 326–331.

[3] A. Roberts, R. Gaizauskas, M. Hepple, et al., The CLEF corpus: semantic annotation of clinical text, in: AMIA Annu Symp Proc, 2007, pp. 625–9.

[4] A. Roberts, R. Gaizauskas, M. Hepple, et al., Building a semantically annotated corpus of clinical texts, J Biomed Inform 42 (2009) 950–966.

[5] B. R. South, D. Mowery, Y. Suo, et al., Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text, J Biomed Inform 50 (2014) 162–172.

[6] S. Yada, A. Joh, R. Tanaka, et al., Towards a versatile medical-annotation guideline feasible without heavy medical knowledge: starting from critical lung diseases, in: Proc of the 12th LREC, 2020, pp. 4565–4572.

[7] T. Rama, P. Brekke, Ø. Nytrø, L. Øvrelid, Iterative development of family history annotation guidelines using a synthetic corpus of clinical text, in: Proc of the 9th Intl Workshop on Health Text Mining and Information Analysis, ACL, 2018, pp. 111–121.

[8] A. Miranda-Escalada, L. Gascó, S. Lima-López, et al., Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources 13390 (2022).

[9] M. Skeppstedt, M. Kvist, G. H. Nilsson, et al., Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text, J Biomed Inform 49 (2014) 148–158.

[10] E. Zhu, Q. Sheng, H. Yang, et al., A unified framework of medical information annotation and extraction for chinese clinical text, Artif Intell Med (2023) 102573.

[11] W. F. Styler IV, S. Bethard, S. Finan, et al., Temporal annotation in the clinical domain, Transactions of the ACL 2 (2014) 143–154.

[12] D. L. Mowery, H. Harkema, W. Chapman, Temporal annotation of clinical text, in: Proc of the Workshop on Current Trends in Biomedical NLP, 2008, pp. 106–107.

[13] M. Marimon, J. Vivaldi, N. Bel Rafecas, Annotation of negation in the IULA Spanish corpus, SemBEaR –Computational Semantics Beyond Events and Roles (2017) 43–52.

[14] Y.-F. Luo, W. Sun, A. Rumshisky, MCN: a comprehensive corpus for medical concept normalization, J Biomed Inform 92 (2019) 103132.

[15] N. Guarino, Formal ontology, conceptual analysis and knowledge representation, Int Journal of Human-computer Studies 43 (1995) 625–640.

[16] B. Smith, M. Ashburner, C. Rosse, et al., The OBO foundry: coordinated evolution of ontologies to support biomedical data integration, Nature biotech 25 (2007) 1251–1255.

[17] S. Schulz, R. Stegwee, C. Chronaki, Standards in healthcare data, in: K. et al. (Ed.), Fundamentals of Clinical Data Science, Springer, Cham(CH), 2019.

[18] J. Millar, The need for a global language–SNOMED CT introduction, Stud Health Technol Inform 225 (2016) 683–685.

[19] A. L. Rector, R. Qamar, T. Marley, Binding ontologies and coding systems to electronic health records and messages, Applied Ontology 4 (2009) 51–69.

[20] E. J. Pacheco, MorphoMap: Mapeamento automático de narrativas clínicas para uma terminologia médica, PhD dissertation. UTFPR, Brazil, 2009.

[21] J. A. Miñarro-Giménez, R. Cornet, M.-C. Jaulent, et al., Quantitative analysis of manual annotation of clinical text samples, Int J Med Inform 123 (2019) 37–48.

[22] M. Ayaz, M. F. Pasha, M. Y. Alzahrani, R. Budiarto, D. Stiawan, The Fast Health Interoperability Resources (FHIR) standard, JMIR Med Inform 9 (2021) e21929.

[23] E. M. Sanfilippo, Ontologies for information entities: State of the art and open challenges, Appl Ontology 16 (2021) 111–135.

[24] J. N. Otte, J. Beverley, A. Ruttenberg, BFO: Basic Formal Ontology, Appl Ontology 17 (2022) 17–43.

[25] S. Gladkoff, L. Han, G. Nenadic, Student's t-distribution: On measuring the inter-rater reliability when the observations are scarce, in: Proc of the RANLP, 2023.

An ontology for mammography screening recommendation

Cindy Acuña[1], Yasmine Anchén[2]*, Edelweis Rohrer[1]* and Regina Motz[1]*

[1] *Facultad de Ingeniería, Universidad de la República, Julio Herrera y Reissig 565, 11300 Montevideo, Uruguay*
[2] *Facultad de Medicina, Universidad de la República, Av. Gral. Flores 2125, 11800 Montevideo, Uruguay.*

Abstract

This article presents the development of a simple Mammography Screening Ontology (MAMO-SCR-Onto) designed to provide mammography application recommendations for the early detection of breast cancer in patients at average risk of the disease. We highlight the simplicity of this ontology that allows, on the one hand, to be a formal communication tool between doctor-patient; on the other, to be an open educational resource for the training of preservice health professionals. The ontology is developed in OWL and accessible with an open licenced.

Keywords

mammography screening recommendation, breast cancer early detection, patients with average risk

1. Introduction

Breast cancer is a significant global health concern, underscoring the vital need for effective early detection methods. The American Cancer Society (ACS), a leading cancer-fighting organization, has established comprehensive guidelines for mammography as a crucial test for the early detection of breast cancer in women with an average risk of contracting breast cancer [1]. In spite of other recognized institutions' recommendations, in the present work we consider ACS's recommendations as a case study without loss of generality. These recommendations are about the frequency application of mammography as a test for early breast cancer detection based on women's values in risk factors. On the other hand, ontologies are widely adopted and proven modelling artefacts for conceptualizing various domains, including education and health. They provide a structured framework to describe vocabularies in terms of concepts (or entity types), instances (or individuals), and roles (or relations), as well as assertions or constraints about them [2]. Notably, the W3C standard ontology language OWL, with its formal semantics based on description logics, facilitates the implementation of a comprehensive set of assertions for a given vocabulary, particularly in the health domain [3]. OWL enables the automatic validation of ontology constraints and facilitates the inference of assertions that

Proceedings International Conference on Biomedical Ontologies 2023, August 28-September 1, 2023, Brasilia, DF, Brazil

*Corresponding authors.

† These authors contributed equally: Yasmine Anchén, Regina Motz, Edelweis Rohrer

cindy.acuna@fing.edu.uy (C. Acuña); yasanmo@gmail.com (Y. Anchén); erohrer@fing.edu.uy (E. Rohrer); rmotz@fing.edu.uy (R. Motz)

0009-0006-4254-4817 (C. Acuña); 0009-0003-8648-2687 (Y. Anchén); 0000-0002-8257-5291 (E. Rohrer); 0000-0002-1426-562X (R. Motz)

CEUR Workshop Proceedings (CEUR-WS.org)

may not have been explicitly declared. This capability empowers the ontology to go beyond the explicitly stated information, uncovering implicit knowledge and facilitating more robust reasoning. Reasoners, i.e. algorithmic implementations such as the tableau algorithm, play a pivotal role in this process. By examining an ontology, reasoners ensure consistency by verifying that the specified constraints are met while drawing logical inferences based on the underlying description logic semantics [4].

The present work makes a significant contribution by modelling and implementing a simple owl ontology designed to provide mammography screening application recommendations for the early detection of breast cancer in women at average risk of the disease. The ontology represents the conditions taken into account to consider a woman within the group "women with an average risk of contracting breast cancer" (group of women to whom the guide is directed), and it also represents specific recommendations that these women receive according to their ages, such as the periodicity of the study. These conditions are modelled as constraints, which the reasoner utilizes to infer the appropriate advice by the guidelines established by the American Cancer Society. This ontology captures the intricate relationships between various conditions that influence the decision to perform screening mammography, such as the risk group to which the woman belongs (taking into account confirmed or suspected genetic mutation, personal history, and exposure to chest radiation treatments) and the woman's age. By formalizing expert knowledge and integrating it into an ontology, we facilitate the development of intelligent decision support systems that can aid healthcare professionals in making informed and personalized screening recommendations. Moreover, this formal specification establishes a standardized framework and is a valuable tool for visualization and reasoning on the conditions. By leveraging the power of ontologies and their reasoning capabilities, we can optimize breast cancer screening protocols and ultimately contribute to improving healthcare outcomes for individuals at average risk.

This paper presents our approach to designing the mammography screening ontology, including selecting relevant concepts, relationships, and axioms. We outline the process of integrating domain-specific knowledge from authoritative sources, such as the ACS recommendations and the NCIt thesaurus [5] into the ontology. Furthermore, we discuss the benefits and challenges of using OWL and reasoning techniques to ensure the ontology's consistency, coherence, and inferential capabilities.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in the field of ontology-based clinical decision support systems. Section 3 presents the ontology's key components and knowledge representation. Section 4 offers the evaluation of the ontology. Finally, Section 5 concludes the paper with a summary of the contributions, limitations, and future directions.

2. Related work

Over the past two decades, ontologies have emerged as crucial artefacts in supporting the study and research of cancer. The adoption of ontologies as a powerful approach has been evident through the development of numerous ontologies in the biomedical domain [6].

One prominent example of a comprehensive ontology in the field of cancer is the National

Cancer Institute Thesaurus and Ontology (NCIt). Designed as an extensive compilation of terminologies, the NCIt covers various facets of cancer research and care [5]. As a fundamental resource, the NCIt plays a pivotal role in facilitating standardized communication and promoting knowledge sharing among researchers, clinicians, and other healthcare professionals in the field of oncology. However, the significance of the NCIt and other ontologies in the cancer domain extends beyond providing a common vocabulary. These ontologies establish a structured framework that enables collaboration, interoperability, and the integration of diverse data sources. They serve as a foundation for knowledge representation, helping researchers and practitioners to capture and organize complex cancer-related information effectively.

In recent years, there has been increasing interest in using ontologies and knowledge graphs in cancer research, as highlighted by the comprehensive review by Silva et al. [7]. This study examined 141 open articles published between 2012 and 2021. Of the papers analyzed, 18 focus on breast cancer-related ontologies. We can observe that breast cancer ontologies serve several purposes, each addressing different aspects of knowledge representation and integration within the field or supporting decision-based systems. Some of these ontologies prioritize the computational accessibility of breast cancer knowledge to represent the information in a format that computer systems can efficiently process [8, 9, 10, 11]. On the other hand, specific ontologies focus on integrating cancer data from various sources, consolidating the information in a unified database [12, 13]. In some cases, the ontologies are designed to facilitate effective communication between doctors and patients, harmonizing terminology and improving mutual understanding [14] or designed to facilitate database user interfaces, where ontology class and relationship labels allow text annotation within the breast cancer domain [15]. In addition, some ontologies in the breast cancer domain claim to represent causal associations between breast cancer incidence and various risk factors, providing valuable insight into the complex relationships involved [16, 8, 17, 18]. Sherimon et al. [19] describe an architecture of six ontologies (patient ontology, breast cancer, symptoms, risk, lab test and questionnaire) to predict breast cancer. However, we cannot find any of these ontologies accessible. Another application involves the semantic modelling of breast cancer-related drugs using ontological approaches, allowing the inference of possible drug repositioning strategies [20]. However, the most notable applications are those that produce clinical inferences [21, 22, 23].

Regarding the radiology domain, we find Sun et al. work presenting an ontology for training breast radiologists [24]. Focusing on the mammography domain, the GIMI Project developed a fundamental mammographic ontology and an ontology for computerized training in breast radiology [25, 26, 27, 28, 29]. Other works on mammographic ontology are Bulubu et al. [30] describing ontology-based mammography annotation and a case-based retrieval approach for breast masses from the digital mammography archive, Ilyass et al. [31] which annotates images in mammograms according to the concepts of a breast cancer ontology which produces RDF metadata and the more recent ones [32, 33, 34, 35].

However, to the best of our knowledge, no open ontologies provide mammography application recommendations for the early detection of breast cancer in women at average risk of the disease as the simple one that we present in this work.

3. Modelling the mammography screening recommendation ontology

ACS recommendations were based on the quality of the evidence obtained and judgment on the balance of benefits and harms of conducting the study, classifying them into strong recommendations (benefits of conducting the study outweigh the desirable effects) or qualified recommendations (there is clear evidence of benefit but less certainty about the balance of harm/benefit, or on the values and preferences of the patient, allows discussion between doctor and patient to make the decision). Based on the age of women at average risk of breast cancer, the ACS recommends Women at average risk of breast cancer should have screening mammography starting at age 40, following the age range recommendations below.

- Women should begin annual screening between the ages of 40 and 44 as a qualified recommendation.
- Women ages 45 to 54 should have screening mammograms annually as a strong recommendation.
- Women age 55 and older should have screening mammograms biennially or have the opportunity to continue screening annually as a qualified recommendation.

Given the relevance of the ACS recommendations, both for the doctor and the patient, the contribution of the ontology described in this section is twofold: (i) conceptualizing the concrete domain by describing those properties or attributes of women which are relevant inputs for applying the recommendation as well as the rules behind it, and (ii) computing a customized recommendation for a set of women instances, which represents a useful tool that aims most doctors and patients quick access to the guide recommendation and to meet the ultimate goal of the earliest prevention of the breast cancer.

Figure 1 depicts the mammography screening recommendation ontology, which has three main classes: *Woman*, *History* and *Recommendation.*

Even though for the scenario described in the present paper, it is important to represent women with an intermediate risk of contracting breast cancer, this group of women is represented by the subclass *With intermediate risk*, of a more general class *Woman*, to facilitate a possible extension of the ontology if maybe in the future recommendations for other groups of women were conceptualized (for example, with high or low risk of breast cancer). Classes *Age Range 40 to 44*, *Age Range 45 to 54*, and *Age gt or eq 55* represent subsets of women with an intermediate risk that have corresponding ages.

As for the class *Woman*, the class *History* is modelled as a general class even though in the present work, we focus on the subclass *Personal Medical History* that represents the set of medical records of women. Properties *hasHistory* and *doesNotHaveHistory* connect each woman to her medical records. The property *hasHistory* connects women to diseases, clinical diagnoses or laboratory test results that they have. The property *doesNotHaveHistory* connects women to those diseases, clinical diagnoses or important laboratory test results so that the doctor knows that the patient does not have it. It is highly relevant that this be made explicit when evaluating the type of breast cancer risk corresponding to the woman and to be able to make a correct recommendation. The ACS guide clarifies that the recommendations are aimed at women with an average risk of breast cancer; these are those without a personal history of breast cancer,

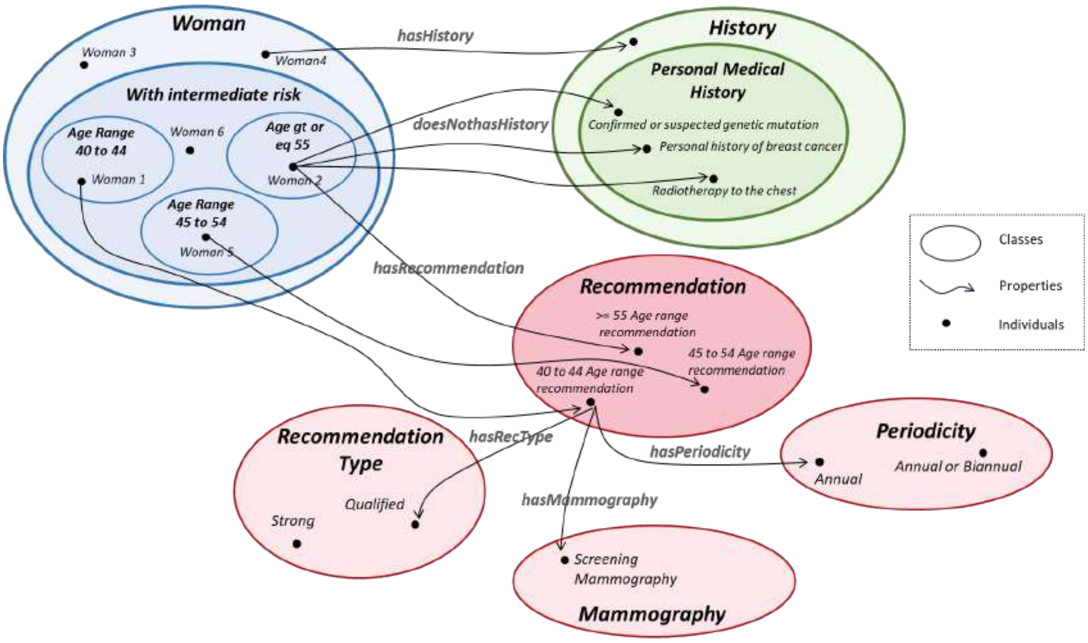


Figure 1: Mammography screening recommendation ontology model.

without suspicion or confirmation of a known genetic mutation that increases the risk of breast cancer and without previous history of chest radiotherapy at a young age. These conditions are represented in our ontology by individuals *confirmed or suspected genetic mutation*, *personal history of breast cancer*, *radiotherapy to chest*, as instances of the class *Personal Medical History*.

The class *Recommendation* represents the set of recommendations provided by the guide. Each particular recommendation has three characteristics or dimensions that describe it: (i) the strength of the recommendation given by the level of evidence to issue the recommendation, represented by the class *Recommendation Type*, which has instances *Strong* and *Qualified* (ii) the recommended mammography study, represented by the class *Mammography* and the instance *Screening Mammography*[1], and (iii) the frequency with which mammography should be performed, represented by the class *Periodicity*, with two instances: *Annual* and *Annual or Biaannual*. Properties *hasRecType*, *hasMammography* and *hasPeriodicity* connect each recommendation instance to the three dimensions.

Restrictions on classes and properties described above are represented in Table 1 as six groups of Tbox and Rbox axioms. Group (1) represents that classes *Woman*, *History*, *Recommendation*, *Recommendation Type*, *Mammography* and *Periodicity* cannot share instances with each other. The axiom (2) defines the class *With intermediate risk* as the set of women that meet the three conditions given by the guide to classify them in the category of women that have an average risk of breast cancer. Group (3) defines subclasses *Age Range 40 to 44*, *Age Range 45 to 54* and *Age gt or eq 55* as subsets of women with an average risk of breast cancer, based on the age range, represented in the implemented OWL ontology by the data property *age*. This is why

[1] Note that for the scenario described in this work, only screening mammographies are considered.

Table 1
Mammography screening recommendation domain Tbox and Rbox.

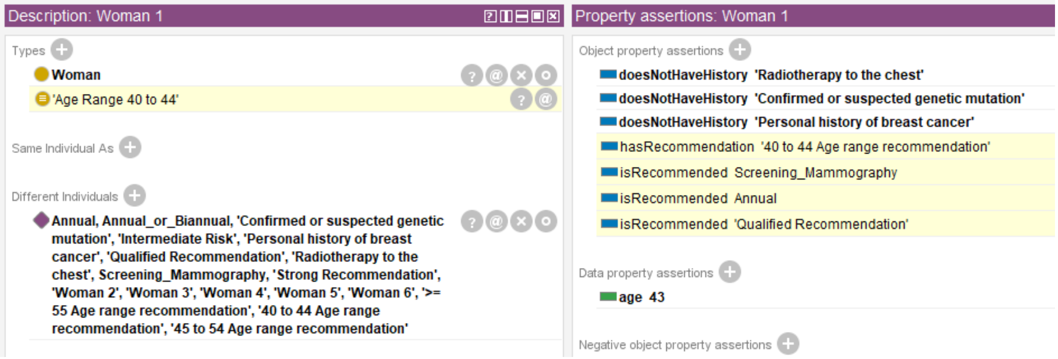
Axiom	Description
$Woman \sqcap History \sqsubseteq \bot$ $Woman \sqcap Recommendation \sqsubseteq \bot$ $Woman \sqcap Mammography \sqsubseteq \bot$ $Woman \sqcap RecommendationType \sqsubseteq \bot$ $Woman \sqcap Periodicity \sqsubseteq \bot$ $History \sqcap Recommendation \sqsubseteq \bot$ (1) $History \sqcap Mammography \sqsubseteq \bot$ $History \sqcap RecommendationType \sqsubseteq \bot$ $History \sqcap Periodicity \sqsubseteq \bot$ $Recommendation \sqcap Mammography \sqsubseteq \bot$ $Recommendation \sqcap RecommendationType \sqsubseteq \bot$ $Recommendation \sqcap Periodicity \sqsubseteq \bot$ $Mammography \sqcap RecommendationType \sqsubseteq \bot$ $Mammography \sqcap Periodicity \sqsubseteq \bot$	Pairwise class disjointness.
$With\ intermediate\ risk \equiv Woman \sqcap$ (2) $\exists doesNotHaveHistory.\{Confirmed\ or\ suspected\ genetic\ mutation\} \sqcap$ $\exists doesNotHaveHistory.\{Personal\ history\ of\ breast\ cancer\} \sqcap$ $\exists doesNotHaveHistory.\{Radiotherapy\ to\ the\ chest\}$	Women with intermediate-risk are those do not meet a set of conditions.
$Age\ gt\ or\ eq\ 55 \equiv With\ intermediate\ risk \sqcap \exists age. \geq 55$ $Age\ gt\ or\ eq\ 55 \sqsubseteq \exists hasRecommendation.\{55\ Age\ range\ recommendation\}$ (3) $Age\ Range\ 40\ to\ 44 \equiv With\ intermediate\ risk \sqcap \exists age. \geq 40 \sqcap \exists age. \leq 44$ $Age\ Range\ 40\ to\ 44 \sqsubseteq \exists hasRecommendation.\{40\ to\ 44\ Age\ range\ recommendation\}$ $Age\ Range\ 45\ to\ 54 \equiv With\ intermediate\ risk \sqcap \exists age. \geq 45 \sqcap \exists age. \leq 54$ $Age\ Range\ 45\ to\ 54 \sqsubseteq \exists hasRecommendation.\{45\ to\ 54\ Age\ range\ recommendation\}$	Each age subclass has corresponding age restriction and recommendation
$\exists hasHistory.\top \sqsubseteq Woman \quad \top \sqsubseteq \forall hasHistory.History$ $\exists doesNotHaveHistory.\top \sqsubseteq Woman \quad \top \sqsubseteq \forall doesNotHaveHistory.History$ $\exists hasRecommendation.\top \sqsubseteq Woman \quad \top \sqsubseteq \forall hasRecommendation.Recommendation$ (4) $\exists hasRecType.\top \sqsubseteq Recommendation \quad \top \sqsubseteq \forall hasRecType.RecommendationType$ $\exists hasMammography.\top \sqsubseteq Recommendation \quad \top \sqsubseteq \forall hasMammography.Mammography$ $\exists hasPeriodicity.\top \sqsubseteq Recommendation \quad \top \sqsubseteq \forall hasPeriodicity.Periodicity$ $\exists isRecommended.\top \sqsubseteq Woman$ $\top \sqsubseteq \forall isRecommended.(Mammography \sqcup Periodicity \sqcup RecommendationType)$	Properties domain and range
(5) Dis(hasHistory, doesNotHaveHistory)	A woman cannot have a clinical situation and does not have it at the same time
$hasRecommendation\ o\ hasRecType \sqsubseteq isRecommended$ (6) $hasRecommendation\ o\ hasMammography \sqsubseteq isRecommended$ $hasRecommendation\ o\ hasPeriodicity \sqsubseteq isRecommended$	A woman is recommended a strong or qualified recommendation, a kind of mammography and a periodicity.

the guide sets a different recommendation for each group of women, which is represented in the ontology as restrictions on instances of classes *Age Range 40 to 44*, *Age Range 45 to 54* and *Age gt or eq 55* that must have the corresponding recommendation.

Axioms (2) and (3) are used by the reasoner to classify instances of the class *Woman* into categories given by descriptions of classes *With intermediate risk*, *Age Range 40 to 44*, *Age Range 45 to 54* and *Age gt or eq 55*, entailing the corresponding recommendation. Group (4) describes restrictions on domains and ranges of properties. (5) and (6) describe restrictions on roles (Rbox axioms). The axiom (5) represents that a woman cannot be connected to the same medical record by both properties *hasHistory* and *doesNotHaveHistory.* In (6), the property *isRecommended* is defined as the superproperty of role chains *hasRecommendation o hasTecType*, *hasRecommendation o hasMammography* and *hasRecommendation o Periodicity*. Besides inferring if the recommendation is for 40 to 44, 45 to 54 or ≥ 55 age range, for each woman, the reasoner directly entails if the recommendation is strong or qualified, the kind of mammography and its periodicity.

4. Validation of the ontology.

The model and restrictions described in the previous section were implemented in the standard W3C ontology language OWL2 by using the Protégé editor. Part of the vocabulary of the implemented ontology was reused from the standard vocabulary of the NCI Thesaurus, e.g. *Woman*, *Mammography* and *History* [36]. The ontology consistency and inferences were verified by using the Hermit reasoner [37].



Description: Woman 1

Types
Woman
'Age Range 40 to 44'

Same Individual As

Different Individuals
Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 2', 'Woman 3', 'Woman 4', 'Woman 5', 'Woman 6', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

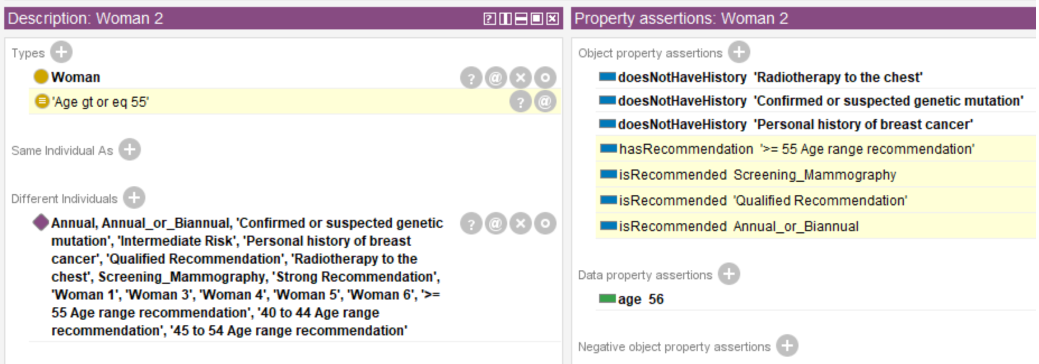
Property assertions: Woman 1

Object property assertions
doesNotHaveHistory 'Radiotherapy to the chest'
doesNotHaveHistory 'Confirmed or suspected genetic mutation'
doesNotHaveHistory 'Personal history of breast cancer'
hasRecommendation '40 to 44 Age range recommendation'
isRecommended Screening_Mammography
isRecommended Annual
isRecommended 'Qualified Recommendation'

Data property assertions
age 43

Negative object property assertions

Figure 2: Reasoner entailments for a woman with intermediate risk and age range 40 to 44.



Description: Woman 2

Types
Woman
'Age gt or eq 55'

Same Individual As

Different Individuals
Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 1', 'Woman 3', 'Woman 4', 'Woman 5', 'Woman 6', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

Property assertions: Woman 2

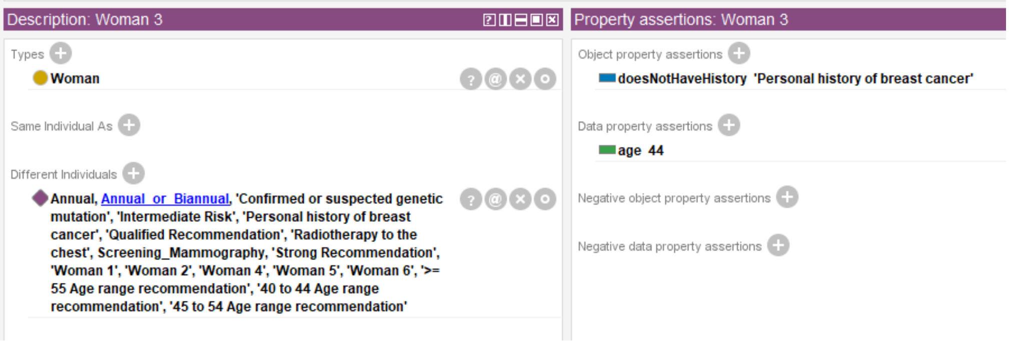
Object property assertions
doesNotHaveHistory 'Radiotherapy to the chest'
doesNotHaveHistory 'Confirmed or suspected genetic mutation'
doesNotHaveHistory 'Personal history of breast cancer'
hasRecommendation '>= 55 Age range recommendation'
isRecommended Screening_Mammography
isRecommended 'Qualified Recommendation'
isRecommended Annual_or_Biannual

Data property assertions
age 56

Negative object property assertions

Figure 3: Reasoner entailments for a woman with intermediate risk and age range ≥ 55.

The ontology was tested with six different instances of the class *Woman*. Three of them meet conditions to be classified as with intermediate risk of breast cancer, each one within a different age range, one woman meets conditions to be classified as with intermediate risk but does not fall into any of the three age range, and the other two women do not meet conditions to be classified as with intermediate risk of breast cancer. Figures 2 to 7 show inferences made by the reasoner for each *Woman* instance. The ontology is available in https://bioportal.bioontology.org/ontologies/MAMO-SCR-ONTO and has the https://creativecommons.org/licenses/by-sa/4.0/ license.



Description: Woman 3

Types

Woman

Same Individual As

Different Individuals

Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 1', 'Woman 2', 'Woman 4', 'Woman 5', 'Woman 6', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

Property assertions: Woman 3

Object property assertions

doesNotHaveHistory 'Personal history of breast cancer'

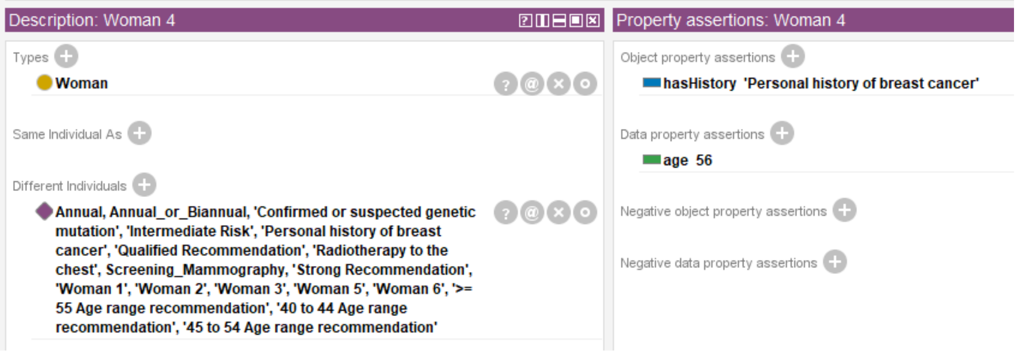
Data property assertions

age 44

Negative object property assertions

Negative data property assertions

Figure 4: Reasoner entailments for a woman that does not meet all conditions for intermediate risk.



Description: Woman 4

Types

Woman

Same Individual As

Different Individuals

Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 1', 'Woman 2', 'Woman 3', 'Woman 5', 'Woman 6', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

Property assertions: Woman 4

Object property assertions

hasHistory 'Personal history of breast cancer'

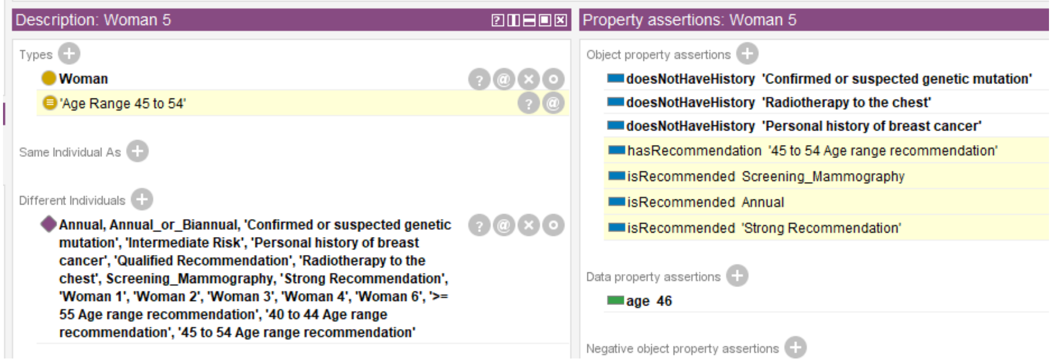
Data property assertions

age 56

Negative object property assertions

Negative data property assertions

Figure 5: Reasoner entailments for a woman that does not meet any condition for intermediate risk.



Description: Woman 5

Types

Woman

'Age Range 45 to 54'

Same Individual As

Different Individuals

Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 1', 'Woman 2', 'Woman 3', 'Woman 4', 'Woman 6', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

Property assertions: Woman 5

Object property assertions

doesNotHaveHistory 'Confirmed or suspected genetic mutation'

doesNotHaveHistory 'Radiotherapy to the chest'

doesNotHaveHistory 'Personal history of breast cancer'

hasRecommendation '45 to 54 Age range recommendation'

isRecommended Screening_Mammography

isRecommended Annual

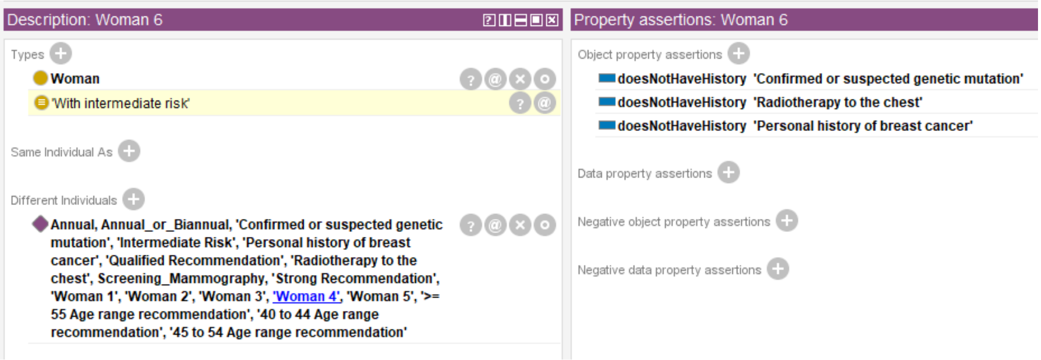
isRecommended 'Strong Recommendation'

Data property assertions

age 46

Negative object property assertions

Figure 6: Recommendation entailed for a woman with intermediate risk and age range 45 to 54.



Description: Woman 6

Types

Woman

'With intermediate risk'

Same Individual As

Different Individuals

Annual, Annual_or_Biannual, 'Confirmed or suspected genetic mutation', 'Intermediate Risk', 'Personal history of breast cancer', 'Qualified Recommendation', 'Radiotherapy to the chest', Screening_Mammography, 'Strong Recommendation', 'Woman 1', 'Woman 2', 'Woman 3', 'Woman 4', 'Woman 5', '>= 55 Age range recommendation', '40 to 44 Age range recommendation', '45 to 54 Age range recommendation'

Property assertions: Woman 6

Object property assertions

doesNotHaveHistory 'Confirmed or suspected genetic mutation'

doesNotHaveHistory 'Radiotherapy to the chest'

doesNotHaveHistory 'Personal history of breast cancer'

Data property assertions

Negative object property assertions

Negative data property assertions

Figure 7: Reasoner entailments for a woman with intermediate risk without the age data.

5. Conclusions and future work

This paper presents a simple Mammography Screening Ontology (MAMO-SCR-Onto) designed to provide mammography application recommendations for the early detection of breast cancer in patients at average risk of the disease. Risk stratification allows for personalized recommendations regarding the frequency, the recommendation strength, and the need for screening mammography tailored to an individual's risk profile. In this sense, MAMO-SCR-Onto is only applied to patients with average disease risk, where the risk is defined based on three conditions of the woman's clinical history. However, the risk could be obtained by applying different evaluation risk models outside the current ontology [38]. This is not necessarily seen as a limitation of the MAMO-SCR-Onto ontology, in the sense that its design aimed to focus the recommendation for exclusively medium-risk patients while maintaining its simplicity to be a helpful tool in doctor-patient communication, and in the training of preservice health professionals.

As a main conclusion, from the point of view of final users, the advantage of using the MAMO-SCR-Onto is that medical professionals and patients can access guideline recommendations in a non-verbose presentation. Notably, the ontology makes it explicit whether these recommendations are classified as strong or qualified, facilitating physician and patient to question the benefits and harms of conducting the study and subsequently assisting decision-making by the patient. Ultimately, this comprehensive ontology enables individuals to make well-informed decisions regarding their healthcare attention.

From the point of view of ontology engineers, another conclusion of our work is regarding ontology maintainability and reusability quality attributes. The MAMO-SCR-Onto represents the conditions that women must meet to receive specific recommendations, established in the ACS guidelines [39]. However, using the ACS guide is only a showcase proposal since it is no difficulty to change ACS guide for any other one [40, 41]. The recommendations in general terms established by the guidelines are concordant, and variations occur when setting the age range in the recommendations by group and the frequency of screening, these constraints are very well identified in the ontology, and they are straightforward the change.

As a future work, the ontology could be extended for women at high risk of breast cancer, representing the recommendations provided for them [42]. More precise information about the corresponding type of risk assessment could be included, such as being based on a statistical model of breast cancer risk assessment, for example, the Tyrer-Cuzick , Clause and Gail models, that the physician at the time of the consultation can assess the risk of breast cancer of the woman with greater precision and in turn provide the relevant recommendations [43, 44, 45].

References

[1] American Cancer Society, https://www.cancer.org/, Last date accessed July 2023.

[2] S. Staab, R. Studer, Handbook on Ontologies, 2nd ed., Springer Publishing Company, Incorporated, 2009.

[3] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. F. Patel-Schneider, U. Sattler, OWL 2: The next step for OWL, J. Web Sem. 6(4) (2008).

[4] P. Hitzler, M. Krötzsch, S. Rudolph, Foundations of Semantic Web Technologies, Chapman & Hall/CRC, 2009.

[5] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, B. Parsia, The National Cancer Institute's Thesaurus and Ontology, Web Semantics 1 (2003).

[6] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, Nucleic Acids Research 39 (2011).

[7] M. C. Silva, P. Eugénio, D. Faria, C. Pesquita, Ontologies and Knowledge Graphs in Oncology Research, Cancers 14 (2022) 1906.

[8] O. N. Oyelade, A. E. Ezugwu, S. A. Adewuyi, Enhancing reasoning through reduction of vagueness using fuzzy OWL-2 for representation of breast cancer ontologies, Neural Computing and Applications 34 (2021).

[9] M. T. D. Melo, V. H. L. Gonçalves, H. D. R. Costa, D. S. Braga, L. B. Gomide, C. S. Alves, L. M. Brasil, OntoMama: An Ontology Applied to Breast Cancer, Studies in Health Technology and Informatics 216 (2015) 1104.

[10] J. Xi, L. Ye, Q. Huang, X. Li, Tolerating data missing in breast cancer diagnosis from clinical ultrasound reports via knowledge graph inference, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Singapore, 2021.

[11] S. Bourougaa-Tria, H. Farah, Semantic ontology for representing breast cancer terminology, in: M. Laouar, V. Balas, B. Lejdel, S. Eom, M. Boudia (Eds.), 12th International Conference on Information Systems and Advanced Technologies "ICISAT 2022", volume 624 of *Lecture Notes in Networks and Systems*, Springer, Cham, 2023.

[12] F. Jusoh, R. Ibrahim, M. S. Othman, N. Omar, Development of breast cancer ontology based on hybrid approach, International Journal of Innovation in Computing 3 (2013) 1.

[13] O. Seneviratne, S. M. Rashid, S. Chari, J. P. McCusker, K. P. Bennett, J. A. Hendler, D. L. McGuinness, Knowledge integration for disease characterization: A breast cancer example, in: Proceedings of the International Semantic Web Conference, Springer, Monterey, CA, USA, 2018.

[14] M. Tapi Nzali, J. Aze, S. Bringay, C. Lavergne, C. Mollevi, T. Optiz, Reconciliation of patient/doctor vocabulary in a structured resource, Health Informatics Journal 25 (2019).

[15] K. Milian, R. Hoekstra, A. Bucur, A. Ten Teije, F. van Harmelen, J. Paulissen, Enhancing reuse of structured eligibility criteria and supporting their relaxation, Journal of Biomedical Informatics 56 (2015).

[16] A. Daowd, M. Barrett, S. Abidi, S. Abidi, Building a knowledge graph representing causal associations between risk factors and incidence of breast cancer, in: Public Health and Informatics, IOS Press, 2021.

[17] J. Bouaud, S. Pelayo, J.-B. Lamy, C. Prebet, C. Ngo, L. Teixeira, G. Guézennec, B. Séroussi, Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project, Artificial Intelligence in Medicine 108 (2020) 101922.

[18] M. Gong, Z. Wang, Y. Liu, H. Zhou, F. Wang, Y. Wang, N. Hong, Toward early diagnosis decision support for breast cancer: Ontology-based semantic interoperability, Journal of Clinical Oncology 37 (2019).

[19] S. P. C, R. Krishnan, M. James, Mellrak: an ontology driven cdss for symptom assessment, risk assessment and disease analysis of breast cancer, in: 2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM), 2021.

[20] Q. Zhu, C. Tao, F. Shen, C. Chute, Exploring the pharmacogenomics knowledge base PharmGKB for repositioning breast cancer drugs by leveraging Web ontology language (OWL) and cheminformatics approaches, in: Pacific Symposium on Biocomputing, 2014.

[21] S. R. Abidi, Ontology-Based Modeling of Breast Cancer Follow-up Clinical Practice Guideline for Providing Clinical Decision Support, in: Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS'07), 2007.

[22] S. Dasmahapatra, D. Dupplaw, B. Hu, P. Lewis, N. Shadbolt, Ontology-mediated distributed decision support for breast cancer, in: Artificial Intelligence in Medicine: 10th Conference on Artificial Intelligence in Medicine, AIME 2005, Aberdeen, UK, July 23-27, 2005. Proceedings 10, Springer Berlin Heidelberg, 2005.

[23] Implementation of an ontological reasoning to support the guideline-based management of primary breast cancer patients in the DESIREE project, Artificial Intelligence in Medicine 108 (2020) 101922. Publisher: Elsevier.

[24] S. Sun, P. Taylor, L. Wilkinson, L. Khoo, An ontology for breast radiologist training, in: Proceedings of the 10th IASTED International Conference on Computers and Advanced Technology in Education (CATE 2007), Beijing, China, 2007.

[25] D. Qi, Development and evaluation of an ontology for a mammographic computer aided diagnosis system, Ph.D. thesis, Aberystwyth, 2006.

[26] S. Sun, P. Taylor, L. Wilkinson, L. Khoo, An ontology to support adaptive training for breast radiologists, in: E. A. Krupinski (Ed.), IWDM 2008, volume 5116 of *LNCS*, Springer, Heidelberg, 2008.

[27] S. Sun, P. Taylor, L. Wilkinson, L. Khoo, Individualised training to address variability of radiologists' performance, in: Proceedings of the SPIE Symposium on Medical Imaging (SPIE-MI 2008), SPIE, San Diego, 2008.

[28] P. Taylor, I. Toujilov, Mammographic knowledge representation in description logic, in: KR4HC 2011, volume 6924 of *Lecture Notes in Computer Science*, Springer, 2012.

[29] GIMI Mammography Ontology, http://sourceforge.net/projects/gimimammography/, Last

date accessed July 2023.

[30] H. Bulu, A. Alpkocak, P. Balci, Ontology-based mammography annotation and case-based retrieval of breast masses, Expert Systems with Applications 39 (2012).

[31] H. Ilyass, F. S. Mohamed, I. Diop, J. Tarik, Ontology-based mammography annotation for breast cancer diagnosis, in: 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015.

[32] J. W. Pereira, M. X. Ribeiro, Semantic Annotation and Classification of Mammography Images Using Ontologies, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2021.

[33] A. Yagahara, Y. Yokooka, G. Jiang, S. Tsuji, A. Fukuda, N. Nishimoto, K. Kurowarabi, K. Ogasawara, Construction of mammographic examination process ontology using bottom–up hierarchical task analysis, Radiological Physics and Tech. 11 (2018).

[34] Y. B. Salem, R. Idoudi, K. Saheb Ettabaa, K. Hamrouni, B. Soleiman, Mammographie image based possibilistic ontological representation, in: 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, 2018.

[35] Y. B. Salem, R. Idoudi, K. Saheb Ettabaa, K. Hamrouni, B. Solaiman, High level mammographic information fusion for real world ontology population, Journal of Digital Information Management 15 (2017).

[36] NCI Thesaurus., https://ncit.nci.nih.gov/, Last date accessed June 2023.

[37] I. Horrocks, B. Motik, Z. Wang, The HermiT OWL Reasoner, in: 1st International Workshop on OWL Reasoner Evaluation (ORE-2012), CEUR-WS.org, 2012.

[38] E. M. Ozanne, B. Drohan, P. Bosinoff, A. Semine, M. Jellinek, C. Cronin, F. Millham, D. Dowd, T. Rourke, C. Block, K. S. Hughes, Which risk model to use? clinical implications of the acs mri screening guidelines, Cancer Epidemiol Biomarkers Prev 22 (2013).

[39] Oeffinger KC, Fontham ETH, E. et al., Breast Cancer Screening for Women at Average Risk:2015 Guideline Update From the American Cancer Society, JAMA. 2015;314(15):1599–1614. doi:10.1001/jama.2015.1278, 2015.

[40] Canadian Task Force Recommendation 2011, https://pubmed.ncbi.nlm.nih.gov/30530611/, Last date accessed July 2023.

[41] Guía práctica clínica de detección temprana del cáncer de mama - Tamizaje y diagnóstico precoz (uruguayan guide 2015), https://www.gub.uy/ministerio-salud-publica/comunicacion/publicaciones/guia-practica-clinica-deteccion-cancer-mama/, Last date accessed July 2023.

[42] D. Saslow, C. Boetes, W. Burke, S. Harms, M. O. Leach, C. D. Lehman, E. Morris, E. Pisano, M. Schnall, S. Sener, R. A. Smith, E. Warner, M. Yaffe, K. S. Andrews, C. A. Russell, American cancer society guidelines for breast screening with mri as an adjunct to mammography, CA: a cancer journal for clinicians 57 (2007).

[43] J. Tyrer, S. W. Duffy, J. Cuzick, A breast cancer prediction model incorporating familial and personal risk factors, Statistics in Medicine 23 (2004).

[44] E. B. Claus, N. Risch, W. D. Thompson, Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction, Cancer 73 (1994).

[45] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, J. J. Mulvihill, Projecting individualized probabilities of developing breast cancer for white females who are being examined annually, J Natl Cancer Inst 81 (1989).

An Application of Natural Language Processing and Ontologies to Electronic Healthcare Records in the Field of Gynecology

Amanda Damasceno de Souza[1], Fernanda Farinelli[2], Eduardo Ribeiro Felipe[3], Armando Sérgio de Aguiar Filho[1] and Mauricio Barcellos Almeida [4]

[1]FUMEC University, Graduate Program in Information and Communication Technology and Knowledge Management (PPGTICGC), Belo Horizonte, MG, Brazil.
[2] University of Brasília (UnB), Brasília, DF, Brazil
[3] Federal University of Itajubá (UNIFEI) Campus Itabira, MG, Brazil
[4] Federal University of Minas Gerais, Belo Horizonte, MG, Brazil

Abstract

Electronic Health Records (EHR) usually comprise medical data sources containing unstructured data. EHRs contain various terms and idiosyncrasies, which prevent reasonable matches to standardized clinical terminologies. That, in turn, impedes information retrieval and the integration of systems of healthcare units, even systems within the same unit. The present article evaluates the application of Natural Language Processing (NLP) to EHR. The research presents a case study examining the connections among the EHR's terms for signs and symptoms, here called the *interface terminology*; a biomedical ontology, here called the *reference terminology*; and the Tenth International Classification of Diseases (ICD-10), here called the *aggregation terminology*. We collected a sample of terms for signs and symptoms in gynecology to test correlations between reference and aggregation terminologies. We report and analyze the main difficulties we encountered during the correlation process regarding the semantics of the terms and the lack of related terms.

Keywords [1]

Electronic health records, clinical terminology, natural language processing, biomedical ontologies.

1. Introduction

Electronic Health Records (EHR) are an essential source of real-world health information for several purposes. Information in EHRs is often recorded in an unstructured format, which poses challenges to using it for computational purposes. Indeed, advances in health information technologies have followed an increasing need for standardized clinical text and terminologies to facilitate information retrieval (IR) and interoperability. Usually, unstructured EHR data have a terminological variety that does not match standardized clinical terminologies, which poses a significant obstacle to achieving IR's objectives [1]. Therefore, an effective means of connecting the ordinary terms found in EHRs with standard medical terminologies could improve IR processes. One option is to map the EHR's terms onto standardized terminologies.

Health terminology standardization is a requirement for achieving effective IR. Structured and controlled data representation is essential when using a terminological system to record medical data. The terminological system consists of techniques and artifacts such as thesauri, controlled vocabularies, taxonomies, and ontologies [2]. Standardized biomedical terminologies are essential because they

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: amanda.dsouza@fumec.br; amandasd81@gmail.com (A. 1); fernanda.farinelli@unb.br (A. 2); eduardo.felipe@unifei.edu.br (A. 3), armando.filho@fumec.br (A. 4). mba@eci.ufmg.br (A. 5) ORCID: 0000-0001-6859-4333 (A. 1); 0000-0003-2338-8872 (A. 2); 0000-0003-1690-2044 (A. 3), 000-0001-5542-7165 (A. 4), 0000-0002-4711-270X (A. 5).

interface clinical data and health care systems, including the EHRs [3]. Standardized terminologies are also valuable resources for enabling interoperability in EHR by collaborating to perform audits, research, benchmarking, and management for hospitals [4].

Our investigation draws on existing literature, such as a study by Schulz et al. [5], who analyze terminology standardization and propose a methodology to connect three types of health terminologies: *interface terminologies*, namely, medical chart text or medical jargon; *reference terminologies*, which are controlled vocabularies and ontologies; and *aggregation terminologies*, which include the International Classification of Diseases (ICD), Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) and others. Our research adopts the denominations employed by Schulz et al. [5]. In this context, research by Rector [6] raises some highly relevant questions.

The gap posed by Schulz et al. [5] requires finding a way to connect the clinical data in an EHR's clinical texts to standardized clinical terminologies, including the ICD, SNOMED-CT, Medical Subject Headings (MeSH), Unified Medical Language Systems (UMLS), and biomedical ontologies such as those found on the OBO Foundry portal. Although Schulz et al.[5] connected three standardized medical terminologies, they didn’t connect any of those to the ordinary language terms found in EHRs. So, significant work remains to be done.

Interoperability among clinical terminologies promotes the generation of innovative products that helps physician better annotate EHRs, contributing to the quality of care and patient well-being. Our research examines a case study about the connections among the terms for signs and symptoms used in the patient’s EHR, a biomedical ontology, and the ICD-10. As its principal contribution, our research verified medical jargon terms that do not correspond to existing biomedical ontologies in the OBO Foundry or OntONeo. As a further contribution, we use OntONeo to connect an EHR’s textual clinical data with the standardized clinical terminologies, which Schulz et al. [5] call *reference terminology*.

2. Methodology

Our interdisciplinary study involves Librarianship and Information Science (LIS), Information Technology, and healthcare fields. We conducted applied research using qualitative, quantitative, and descriptive methods. We followed the tenets of those mentioned above, well-established researchers to standardize biomedical terminologies by adopting three designations: i) interface terminologies, which stand for ordinary language texts recorded in EHRs; ii) reference terminologies, which are ontologies and controlled vocabularies; iii) aggregation terminologies, which are ICD-10 artifacts [5]. Then, we applied natural language processing (NLP) techniques and domain ontologies, specifically OntONeo [7]. Our methodology relied on NLP to extract and analyze signs and symptoms from clinical texts, ultimately connecting them to the standards by mapping them through ontology. We performed the usual pre-processing preparation stages of the free text, including treatment of stop-words, and case-folding techniques, excluding break-lines. In the information-extraction step, we developed specific algorithms to locate signs and symptoms and compare them to a list of signs and symptoms previously prepared by domain experts seeking to improve the automatic task of term identification. The information extraction was performed in a large private hospital, which provided a sample of 32,291 real EHRs containing medical notes in free text. These groups of notes cover the evolution and medical history of patients from the gynecology department during the year 2018, and their use was authorized through the appropriate administrative and ethical processes. [8]

The medical team created a pre-list of signs and symptoms to delimit the algorithm for data processing. Other sources of information used in the pre-list of signs and symptoms were the National Library of Medicine (NLM) Classification 2020 Summer Edition [9], Wikipedia [10,11], Falcão Junior et al. [12], and ICD-10. For the pre-list of signs and symptoms, it was necessary to include data on the following systems: circulatory and respiratory; digestive and abdomen; skin and subcutaneous tissue; nervous and musculoskeletal; and urinary. The pre-list also included terms about cognition, perception, emotional state and behavior, speech and voice, and general signs and symptoms. This pre-list was validated by a gynecologist, i.e., a domain expert.

The next step was determining the most frequent signs and symptoms in the general population and their quantity in the EHRs. This list of signs and symptoms was created in a text file, which was, in turn, read by the algorithm to create a list (array) of terms found. In the database, the result of this reading was segmented according to the type of analysis ("anamnesis" and "evolution"). Therefore, in each record whose information was extracted from the hospital institution, the correspondence between those signs and symptoms (already available in the list in memory) that appeared was traced. A data structure was organized by a pair key, namely, value, called a *dictionary* in Python programming language. This model allows the storage of the ICD code (key) and the identification of its quantity (value). This data structure was later recorded in a spreadsheet format file.

The last step was to check the frequency of the interface terminology and its proper correspondence to the reference terminology. This analysis step was performed by a medical expert specializing in gynecology. After mapping the terminologies, the number of terms present in the interface terminology and reference terminology was quantified to verify the percentage of connectivity (match) between the clinical terminologies. Finally, the results were described for their respective groups.

2.1 Mappings between Terminologies

In mapping the interface terminology onto the reference and aggregation terminologies, the ABNT ISO/TR 12300 standard was taken as the base [13]. The steps for mapping were as follows:

1) Document the mapping process between clinical terminologies (Table 1).

2) Verify the semantic equivalence between terms (Table 1).

3) Utilize a source mapping for terms with multiple synonyms (Table 1).

4) Analyze risk factors and document ways to ensure consistency in mapping.

5) Clarify the meaning and fully use the form for abbreviations in the interface terminology.

6) Map the target terms of the reference terminology selected from Health Science Descriptors (DeCS)[2][14], created by The Latin American and Caribbean Center on Health Sciences Information[3]. Such terminology was developed from Medical Subject Headings (MeSH) [15], and OntONeo as the reference terminology belongs to the OBO-Foundry and aligns with principles of good practices in developing ontologies. Also, map the ICD-10 as the aggregation terminology since this is the classification used in the hospital institution whose data supported this research (Table 2).

7) Create a mapping table to demonstrate the types of interoperability verification: interoperate one term for one, interoperate one term for many terms, interoperate many terms for one term, interoperate many terms for many terms, and do not interoperate (Table 2).

It should be noted that the corpus of unstructured medical data used in the study was created in Portuguese, so the controlled vocabulary used was DeCS. It is a multilingual thesaurus that "[…] to serve as a unique language in indexing articles from scientific journals, books, congress proceedings, technical reports, and other types of materials, as well as for searching and retrieving subjects from scientific literature from information sources available on the Virtual Health Library (VHL) such as LILACS, MEDLINE, and others".[14] DeCS is a translation of MeSH [15] into Portuguese, also

[2] In Portuguese: Descritores em Ciências da Saúde. Available on the internet in: https://decs.bvsalud.org/ Access Jun. 01 2023

[3] In Portuguese: BIREME. Available on the internet in: https://www.paho.org/en/bireme. Access Jun. 01 2023.

providing terms in Spanish and French. Therefore, the research also registered the controlled vocabulary terms in English, i.e., the original version from MeSH, for publication in this language.

Table 1
Preliminary Steps for Mapping Clinical Terminologies

Terminology	Mapping	Terminology	Support (source mapping)
Interface terminology	- Check diagnostic terms, signs and symptoms - Anamnesis/Evolution of Gynecology	Anamnesis and Evolution of Gynecology	-Gynecology Anamnesis Books/ Gynecology and Obstetrics Guidelines-Wikipedia. -Domain expert
Reference terminology	- Check which are and quantity of diagnostic classes, signs and symptoms of Gynecology.	*OntONeo*	-DeCS/MeSH
Aggregation terminology	-Check which are and quantity of classifications for diagnosis, signs and symptoms of Gynecology.	International Classification of Diseases - ICD-10	-Domain expert

Fonte: [8].

Table 2
Mapping of Terms

Mapping	Relation	Final decision
Interoperate one term for one	A single source class is linked to a single target class or term	Retain
Interoperate one term for many terms	A single source class is linked to multiple target classes or terms	Define a class according to basic formal ontology (BFO) and choose term that poses no clinical risk
Interoperate many terms for one term	Multiple source classes are linked to a single target class or term	Define a class according to BFO and choose term that poses no clinical risk
Interoperate many terms for many terms	Multiple source classes are linked to multiple target classes or terms	Define a class according to BFO and choose a term that poses no clinical risk

Source: [8], [16].

3. Results

The first part of the results presents the frequency of terms found in the free-text fields of the EHR. We retrieved approximately 80 types of signs and symptoms in addition to stop-words, abbreviations, and negation expressions, which revealed the complex challenges of planning any automatic initiative. (Table 3). The principal signs and symptoms found refer to frequent complaints in gynecology: pain (n=3671); bleeding (n=2889); edema (n=800); pruritus (n=757); and discharge (n=664).

Table 3
Examples of Signs and Symptoms in Interface Terminology

Terms	Absolute Frequency (n)
Pain	3671
Bleeding	2889
Edema	800
Itching	757
Discharge	664
Dysmenorrhea	456
Vomiting	398
Nausea	336
Abdominal pain	318
Fever	308
Nausea	305
Pelvic pain	298
Tension	219
Metrorrhagia	182
Abnormal uterine bleeding	169
Heartburn	165
Atrophy	163
Headache	154
Coma	147
Depression	133
Urinary incontinence	132
Anxiety	122
Vomiting	119
Pelvic pain	110

Source: [8].

For interface terminologies, we surveyed DeCS[14] to check definitions and synonyms, following methodological step 3 (use a source mapping for terms with multiple synonyms). Then, we compared the correlated terms found with both tables of signs and symptoms of ICD-10 [17] and OntONeo [7]. By methodological steps 6 and 7, we then mapped the target terms of the reference terminology (selected from DeCS/MeSH and OntONeo as the reference terminology [...]) and created a mapping table to demonstrate the types of interoperability verification[...]) displayed in Table 1; the results are presented in Table 4.

Selected examples demonstrate the correspondence between the clinical terminologies. We verified that for signs and symptoms frequently reported in gynecological consultations, there was no correspondence between the term from the interface terminology, e.g., “itching,” and that in the reference terminologies. Another example of signs and symptoms frequently reported in gynecological consultations, there was no correspondence between the term from the interface terminology, e.g., “Irregular menstrual cycle,” and that in the aggregation terminologies.

The term was present only in the DeCS/MeSH-controlled vocabulary. The term “irregular menstrual cycle” did not match the clustering terminology. Only the term “dysmenorrhea” found a match in the three types of clinical terminologies, i.e., interface (EHRs); reference (OntONeo and DeCS/MeSH); and aggregation (ICD-10). Table 4 shows no correspondence between the EHRs’ terms and ICD-10; similarly, the EHRs’ terms did not correspond to OntONeo. The interface terminology terms that were not matched in the reference terminology, OntONeo, will be added to this ontology. Language variations will be added to the ontology’s enrichment, specifically in synonyms.

Table 4
Examples of correlated terms found compared with signs and symptoms of OntoNeo, DeCS/MeSH, and ICD-10 [8].

EHRs	OntONeo	DeCS/MeSH	ICD-10
Irregular menstrual cycle	Process - biological_process - reproductive process - single organism reproductive process - ovulation cycle - menstrual cycle - Quality - Phenotypic abnormality - Abnormal genital system morphology - Abnormality of the menstrual cycle	Menstrual cycle	–
Itching	–	Pruritus	L29.0 Pruritus ani L29.2 Pruritus vulvae L29.3 Anogenital pruritus, unspecified L29.8 Other pruritus L29.9 Pruritus, unspecified Itch NOS
Dysmenorrhea	- Quality - information carrier-sintoma - nervous system symptom - sensation perception - pain	Dysmenorrhea	R10 Abdominal and pelvic pain R10.1 Pain localized to upper abdomen
Painful urination	- Quality - information carrier-sintoma - nervous system symptom - sensation perception - pain - renal colic	–	R30 Pain associated with micturition

Source: [8].
Note: The dash (–) signifies the absence of terms.

The second part of the results reports the mapping among the terms. As seen in Table 5, when applying the mapping according to the ABNT ISO/TR 12300 Standard [13], between interface terminology for reference terminology (OntONeo), 60.15% (n=80) of the signs and symptoms do not interoperate. The second most frequent mapping type was *interoperated one term for one term*.

Table 5
Mapping Interface Terminology Terms to the Reference Terminology (OntONeo)

	Signs and Symptoms	
Interoperability	**n**	**%**
Interoperate one term for one	27	20,30
Interoperate one term for many terms	5	3,76
Interoperate many terms for one term	18	13,53
Interoperate many terms for many terms	3	2,26
Non-interoperable	**80**	**60,15**
Total	133	100

Source: (8).

In Table 6, when applying the mapping according to the ABNT ISO/TR 12300 Standard [13], between interface terminology to aggregation terminology (ICD-10), it can be seen that 53.15 % (n=76) of the signs and symptoms do not interoperate.

Table 6
Mapping Interface Terminology Terms to Aggregation Terminology (ICD)

Interoperability	**Signs and Symptoms**	
	n	**%**
Interoperate one term for one	43	30,07
Interoperate one term for many terms	13	9,09
Interoperate many terms for one term	6	4,20
Interoperate many terms for many terms	5	3,50
Non-interoperable	**76**	**53,15**
Total	**143**	**100**

Source: [8].

4. Discussion

Some aspects of the results presented so far are worth stressing and discussing. For example, Table 3 indicated that the term "irregular menstrual cycle" is correlated to the OntoNeo Ontology and DeCS/MeSH terms but did not show a corresponding term in the ICD-10. The term "itching" is absent in the ontology. "Dysmenorrhea" is already included in the three terminologies. The last example, "painful urination," appears in the ontology and the ICD-10. Table 2 shows the semantic variety to represent signs and symptoms in clinical terminology and the absence of terms in these instruments. Applying the matching between interface terminology and the reference terminology (OntONeo) indicates that 60.15% of the signs and symptoms do not interoperate.

In matching terms in the interface terminology to those in the reference terminology for OntONeo classes, we mapped multiple interface terminology source classes to multiple classes or target terms in the ontology. Defining a single class according to the BFO was necessary to avoid multiple inheritances. We performed the same procedure for a single source class in the interface terminology, which we mapped to multiple classes or target terms in the reference terminology (OntONeo ontology). In the case of multiple interface terminology source classes, we mapped to a single ontology target class or term. The excess terms were used to enrich the OntONeo synonym class.

In mapping terms from the interface terminology to terms in the aggregation terminology (ICD), we found that the type "does not interoperate" stood out, and signs and symptoms were absent in 53.15% (Table 6). It is worth noting that the mapping of "interoperates many terms for many terms" obtained an equivalence of 3.50% of the signs and symptoms. A significant absence of interface terms was detected in the aggregation terminology (ICD-10), demonstrating the need to review and update this artifact for better application in the medical profession's clinical practice.

Schulz et al. [5] note the difficulty in reconciling interface terminologies, reference terminologies (e.g., SNOMED CT), and aggregation terminologies (e.g., ICD-11), tying that difficulty to the distinct functions of each terminology. Such difficulties were demonstrated in this research through the percentages of terms that did not interoperate with each other in clinical terminologies: 60.15% of signs and symptoms between interface terminology and reference terminology (OntONeo), and 53.15% of signs and symptoms did not interoperate in the mapping step between interface terminology and aggregation terminology (ICD-10).

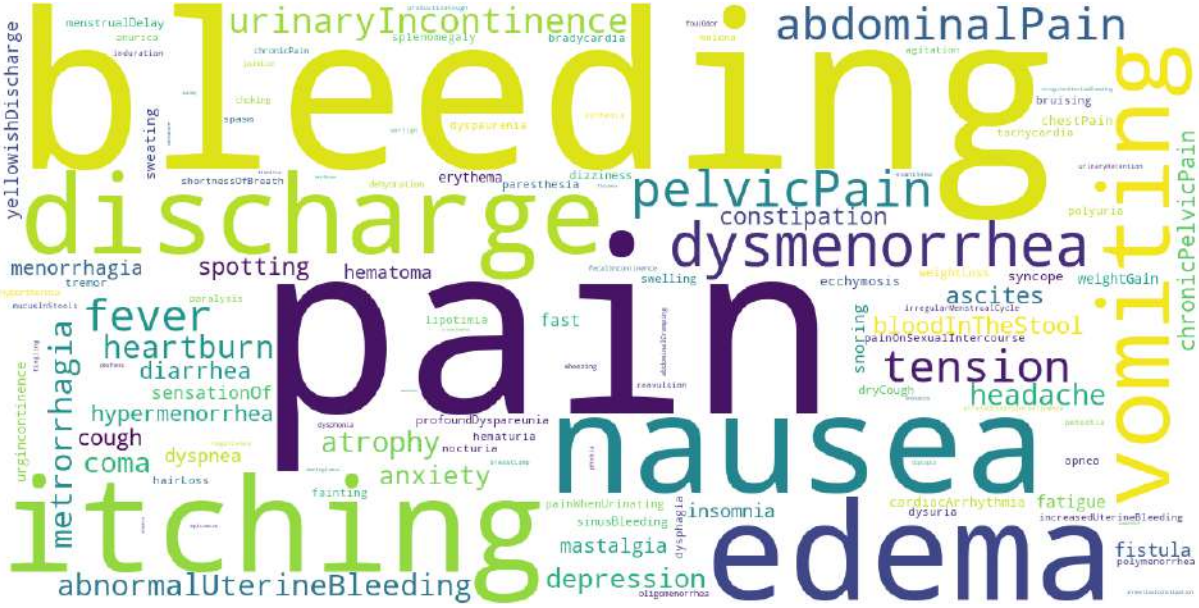


Figure 1: Word Cloud of Most Frequent Signs and Symptoms.
Source: Souza [8].

The frequencies or percentages between mappings indicate that interface terminology is more distant from reference terminology than aggregation terminology. This is explained by physicians' greater familiarity with the aggregation terminology than with the reference terminology; consequently, the terms used in reporting the open fields of the EHR resemble the terms in ICD more than those in OntONeo[7]. Terms in the interface terminology tended to be absent from the aggregation and reference terminologies, demonstrating that interface terminology has richly diverse terms. Notably, the sample used in this research was satisfactory; the richness of its terminology, as shown in Figure 1, enabled it to contribute substantially to the OntONeo ontology and other biomedical ontologies.

5. Final Considerations[4]

Having modified the second step of the proposal by Schulz et al. [5], we performed the connections (mappings) for this research in two steps: first, we mapped interface terminologies to reference terminologies, and subsequently, we mapped the interface terminologies to aggregation terminologies. Instead of the reconciliation step between reference and aggregation terminologies, we mapped interface terminologies to aggregation terminologies. This modification was necessary because we focused on analyzing the mappings between interface terminology and clinical terminologies.

The medical jargon (interface terminology) used in clinical practice proved to be different and distant from standardized terminologies such as ontologies (reference terminologies) and even from ICD-10 (aggregation terminology). This research described some differences in syntax and semantics that posed obstacles to achieving interoperability between information health systems. To reduce these differences, we propose using existing knowledge representation resources in the information science field and the assistance of clinical librarians.

We identified several issues with spelling, punctuation, and typographical errors in the analyzed text. We realized the difficulties in applying NLP techniques to real-world texts and foresaw that ontology could reduce the peculiarity of human notes, helping to achieve the goal of harmonization. As

[4] **Funding:** No grant supported this research
Declarations Ethics approval: The study was approved by the local institutional review board (CAAE:03384418.0.0000.51259).
Competing interests: The authors declare that they have no competing interests.

an additional contribution, we created a computational lexicon (corpus in healthcare) in Portuguese, which can help create algorithms for the domain of gynecology.

One of the main aspects explored in the research was the issue of semantics and syntax of the terms. In this, we aimed to address a primary difficulty in analyzing the medical jargon used in interface terminology, namely, its epistemological aspects, which depend heavily on the medical context. Thus, ontology is an artifact that should be used in seeking a solution to this difficulty.

6. References

[1] S. W. Smith, R. Koppel. Healthcare information technology's relativity problems: A typology of how patients' physical reality, clinicians' mental models, and healthcare information technology differ. J Am Med Inform Assoc. 21(2014):117-31. doi: 10.1136/amiajnl-2012-001419.

[2] N. F. de Keizer, A. Abu-Hanna, J.H. Zwetsloot-Schonk. Understanding terminological systems. I: Terminology and typology. Methods Inf Med. 39 (2000):16-21.

[3] J. Rogers. Using Medical Terminologies. (2005). Available from: http://www.cs.man.ac.uk/~jeremy/HealthInf/RCSEd/terminology using. Htm. Accessed on: 05 Mar. 2019.

[4] J. A. Miñarro-Giménez, R. Cornet, M. C. Jaulent, H. Dewenter, S. Thun, K. R. Gøeg, D. Karlsson, and S. Schulz. Quantitative analysis of manual annotation of clinical text samples. Int J Med Inform. 123 (2019):37-48. doi: 10.1016/j.ijmedinf.2018.12.011.

[5] S. Schulz, J. M. Rodrigues, A. Rector, C. G. Chute. Interface Terminologies, Reference Terminologies, and Aggregation Terminologies: A Strategy for Better Integration. Stud Health Technol Inform, 245(2017):940-944.

[6] A. L. Rector. Clinical Terminology: Why is it so Hard? Methods of Information in Medicine, Stuttgart, 38:147-157, 1999.

[7] F. Farinelli. et al. OntONeo: The Obstetric and Neonatal Ontology. In: Conference: International Conference on Biomedical Ontology 2016, ICBO, Corvallis, Oregon, USA, August 2016. Available at: https://www.researchgate.net/publication/304254064_OntONeo_The_Obstetric_and_Neonatal_Ontology.

[8] A. D. Souza. Clinical Practice Discourse and Standardization Terminologies: Investigating the Connection. Pós-Graduação em Gestão e Organização do Conhecimento, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte. 2021. [Portuguese]. Available at: http://hdl.handle.net/1843/38044. Accessed on: 03 Jul. 2023.

[9] S. R. Willis. NLM Classification 2019 Summer Edition Now Available. The NLM Technical Bulletin, Bethesda, n.430, e4, Sep-Oct 2019. Available at: https://www.nlm.nih.gov/class/index.html. Accessed on: Jul. 21, 2020.

[10] Wikipédia, a enciclopédia livre. Lista de sintomas médicos. Esta página foi editada pela última vez às 04h13min de 13 de janeiro de 2019b. [Portuguese].Available at: https://pt.wikipedia.org/wiki/Lista_de_sintomas_m%C3%A9dicos.

[11] Wikipédia, a enciclopédia livre. Sinal médico. Esta página foi editada pela última vez às 18h56min de 17 de agosto de 2018.[Portuguese] Available at: https://pt.wikipedia.org/wiki/Sinal_m%C3%A9dico.

[12] J. O. A. Falcão Júnior *et al.* Ginecologia e obstetrícia: assistência primária e saúde da família. Rio de Janeiro: MedBook, 2017.[Portuguese]

[13] Associação Brasileira de Normas Técnicas. Relatório técnico ISO/TR 12300: Informática em saúde – princípios de mapeamento entre sistemas terminológicos. 28.11.2016. Rio de Janeiro: ABNT, 2016.[Portuguese]

[14] Health Sciences Descriptors: DeCS [Internet]. 2017 ed. São Paulo (SP): BIREME / PAHO / WHO. 2017 [updated 2017 May; cited 2017 Jun 13]. Available at: http://decs.bvsalud.org/I/homepagei.htm

[15] National Center for Biotechnology Information, U.S. National Library of Medicine. (2022). MeSH (Medical Subject Headings). Bethesda: NLM. Available at: https://www.ncbi.nlm.nih.gov/mesh/.

[16] R. Arp, B. Smith, and A. D. Spear. Building Ontologies with Basic Formal Ontology. Cambridge, MA: The MIT Press, 2015.

[17] ICD-10. (2022). Symptoms and signs involving the genitourinary system R30-R39. Available at: https://www.icd10data.com/ICD10CM/Codes/R00-R99/R30-R39.

The potential of ontologies for the empirical assessment of machine learning techniques in operational oceanography

Enrique Wulff [1]

[1] *Marine Sciences Institute of Andalusia, Spanish National Research Council (CSIC) Campus del Río San Pedro Cadiz 11510, Spain*

Abstract

A role for ontologies is key for the digital transformation of operational oceanography processes to the adoption of artificial intelligence and machine learning. Marine ontologies, a common concept among these tools, can lead to lower costs and more flexibility in identifying and classifying marine data. This study explores a demonstration that proves the potential of ontologies to fulfill the requirements outlined in the case of how to visualize computer datasets. A selective network of records, including visual and textual features that can be annotated from video and image sequences, with subsea parameters as the target of interest. The sample is divided into ontology and machine learning (ML) datasets to predict the importance of data visualization methods. The predicted suitability is strong with data classification that belongs to the machine learning dataset. However, the initial results from the study are encouraging, because ontologies' tools are proposed as automatic reasoning mechanisms. This proof of principle shows that it is almost guaranteed that marine ontologies can be built to make visual patterns for marine data usable by different communities, which could be used to identify "interesting" functions at the intersection of computer vision and machine learning in general.

Keywords

Ontology, machine learning, artificial intelligence, data visualization, classification

1. Introduction

An improved ontological representation of marine data as a paradigm for pattern analysis software development requires more work on combining different modes of inference (OWL, ML), the design of algorithms for data classification (DC) and visual data recognition (DR) for signal and image analysis [1]. This poses the problem of how should marine databases be represented. An ontology of a domain is an "explicit formal specification of the terms in the domain and relations among them" [2]. An ontology fully describes the subject area as a dictionary, in a way it is the ideal tool when we focus on the generation of contextual descriptions for images (in 3D shape retrieval for example [3]). Most of pattern analysis algorithms in oceanography, are to be used for object detection and recognition research, motivated by this challenge it can be proved that an ontology could be a relevant approach to the problem of marine data recognition and classification.

The marine data received from wireless sensor networks are heterogeneous in nature. For instance, the existing marine acoustic data cannot meet the amount of data required for training models [4]. In particular, positioning and orientation systems, and other sensor technology, is based on multi-beam echo sounder system acceptance and quality assurance. An automated system producing multiple overlapping range images that was the first for correctly registered mapping of the ocean floor [5]. Whether data come from GIS technology, the Web or any other present or future approach they share common ground [6]. A role for ontologies is key in the development of application software for the acquisition, analysis and display of real time marine data, for the generation of model scenario databases for their retrieval, and display at the time of an event and for the decision support systems following a

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: enrique.wulff@icman.csic.es
ORCID: 0000-0001-8104-6147

standard procedure [7-8]. To define and develop intelligent systems, has been proposed in recent times, giving a rise in both precision and recall as well as facilitating system interoperability through data harmonization [9-12]. Ensuring interoperability between marine databases is a huge challenge. Terms and codes used to structure exploit the data comes from many sources and are continuously evolving.

The problem pattern analysis (PA) is facing consists in finding an adequate visualization, a "good" figure, since humans are only capable of perceiving objects in at most three dimensions [13]. This means that pattern analysis has to find a method to reduce the heterogeneity of the set of data under study, thus allowing an analysis of the problem of stability of pattern. For practical reasons usually only recognition and classification of those data are allowed (best practices must be carried out by focusing on structure and naming consistency) [14]. Image recognition tasks are at the centre of the ongoing machine learning revolution, an approach that in the monitoring of coastal seas is focused on using automated classification algorithms based on random forest or deep learning approaches [15]. However, the field of marine image processing lacks the large numbers of annotations in images required [16]. The lack of correspondence between the visual representation of the image and its meaning calls for the performance of Machine Learning, expressed through semantic resources such as ontologies.

The problem of trying to solve the visual parameters of images or videos focuses on tasks such as object detection, data recognition, and multi-level data classification. Such an example could be that of studying how the air and sea interact with each other during El Niño/La Niña onsets, by using pattern analysis with ocean data assimilation techniques [17]. This is an issue where content-based image retrieval is approached in terms of Machine Intelligence [18][19]. As such Pattern Analysis and Machine Intelligence (PAMI) is an element of scholarship proposed in the last thirty years and where it has been a continuous need to develop new data recognition and classification methods and advanced equipment for solving modern practical problems [20].

1.1. Pattern Analysis [and Machine Intelligence (PAMI)] and Marine Ontologies

Rethinking pattern analysis of marine data means to investigate the rich variety of application scenarios offered by marine ontologies. While otherwise adding value to public data using semantic web axioms and machine learning to support annotation contribute to pose and solve issues involving ocean data classification.

Application of ontologies in ocean data grows out of an Artificial Intelligence (AI) engagement with marine data metrics of interoperability and reuse. Ontologies serve as such a tool and method to assess the added value robotic technology brings into the marine environment (autonomous underwater vehicles (AUVs) or (ocean floor observation systems) OFOSs). From a pattern recognition point of view, ontologies for describing sensors and sensor networks work in the context of Sensor Web applications. Knowledge representation in the Internet of Things (IoT) presents a general architecture of Sensor Web applications. And that is why it provides huge numbers of interconnected data across an extended variety of various ocean regions, which classifications depend on the specific context and resources of LinkedData.

By using ontological representation, the best of technical progress, undertaken by a community to unambiguously set definitions and interconnect concepts in various field, is captured. The use of ontologies for representing database entities has proven to be advantageous in the field of Pattern Analysis and Machine Intelligence (PAMI) (see Table 1).

Table 1
The main features provided by ontologies in support of PAMI

Ontology feature	Utility in PAMI
Classes and relations	When ontology reasoning is applied to sensor data, rdf:type will be connected to a class name of an ontology
Domain vocabulary	Ontologies provide a domain vocabulary that can be exploited to create a dense network of relationships among the entities, and serve software applications, and GIS
Metadata and descriptions	Biodiversity data, especially in marine domain, have database entities represented as ontologies where these last are primarily used for metadata that describe raw data providing contextual information
Axioms and formal declarations	Ontology axioms and applied reasoning on them are related to the recognition of object presence in a time interval

The concept of marine ontologies may be the solution in developing systems and workflows that would meet the various possible marine data requirements and from them derive up to standard products/maps without human assistance except at the user interface. As shown in Figure 1, research on ontology topics can be followed from different perspectives. The index is the percentage of the publications in the ontology sub-areas of research. It covers semantic web, web services and so forth. Especially, the semantic web, data integration, and web service have attracted the attention of a large number of researchers in recent years while the research on the topics of data source, relation extraction and heterogeneous data seems less consistent. One element is the major cause of these problems, as far as a common ontology for marine data is necessary to enable exchange and integration of data. Terminology is used to describe similar data can vary between marine specialties or world ocean regions, which can complicate data searches and data integration.

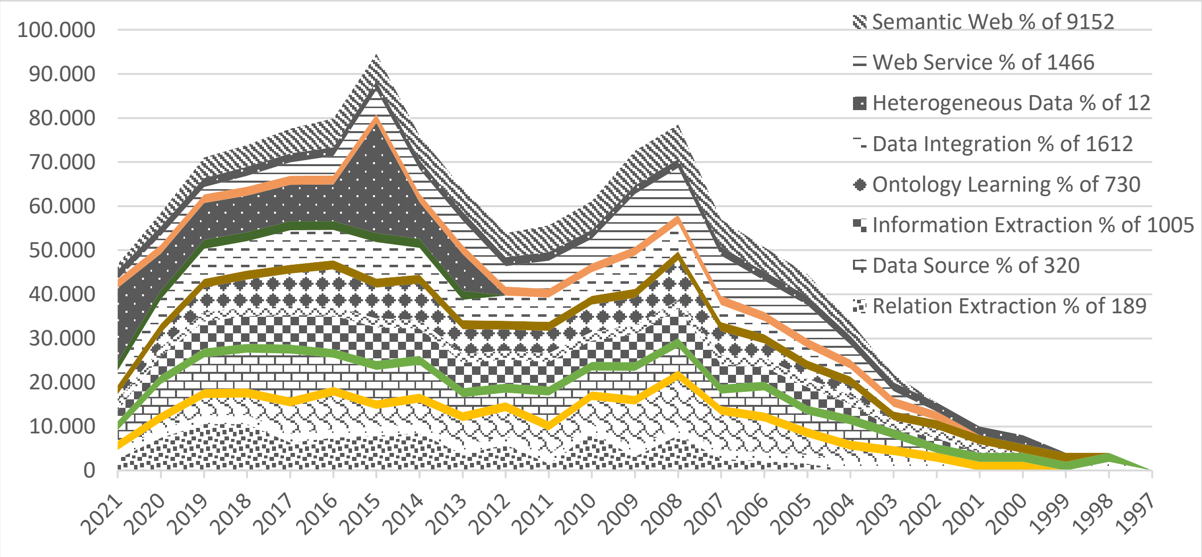


Figure 1: Ontology subareas of research (dependent variable: percent of publication in ontology sub-areas)

The ontology-based research illustrates especially how those involved with marine data should be informed about marine ontology developments. Opportunities to enhance their development will contribute to the success of ontology in the way that certain concepts and ideas start to unfold. It is customary to consider emerging observational patterns making sense out of methods that captures concepts leading to finding out what is visible. To conduct this insight, for example in coastal web atlases (CWA), developers should intensify future efforts to improve data discovery, sharing, and integration on the base of ontologies.

1.2. Ontologies and Marine Robotics

Marine data classification has been studied widely in the field of marine robotics; while pattern recognition is a process of finding regularities and similarities in data using machine learning data which is the perspective of marine robotics. Marine robotics has undergone a phase of dramatic increase and its quantitative landscape, status quo and current workflow is shaped by its own pattern analysis, data recognition and classification issues.

From the point of view of marine robotics key issues in ocean data management concern two different PAMI realities representing, detecting, and tracking features and the process of integrating real sensor data with a model of an ocean process. As a standard knowledge representation ontology can facilitate the development of these marine robotic applications in various ways:

- Providing a consistent set of terminology (domain vocabulary), and concepts in the robot's knowledge representation (definitions, relations, domain axioms and taxonomy)
- Enabling design pattern guidelines for content analysis of complex tasks, environment, etc.
- Ensuring a common repository of knowledge that can be shared among various robotic systems
- Highlighting more efficient new relations through the analysis of data generated using ontologies

1.3. Contribution of this paper

The purpose of this paper is to identify relevant pattern analysis research in marine data classification and recognition, and to review its intersection with the state-of-the art in marine ontologies. It focuses on the 3D modeling and analysis domain, computer vision and interactions are described for machine learning (ML) and marine ontologies.

2. Method

All the R&D efforts in pattern analysis, classification and recognition of data have been kept rising over the current period (1991 to 2021). To obtain a general understanding of this research question concerning marine data we systematically reviewed the IEEE Pattern Analysis and Machine Intelligence, IEEE Access, IEEE Journal of Oceanic Engineering, Sensors, and Information Visualization. Initially, we identified the appropriate subset of articles from these conferences and journals. We then conducted an in-depth qualitative analysis of the relevant work, re-removing and refining the characteristics of the marine data interaction of PA. The histogram theory inspired us to take a general approach to this analysis, which systematically analyzes the data until significant categories appear. This methodological approach is based on define and refine categories based on a representative set of qualitative data, here are documents that are then used to progressively build a theoretical model. This approach has been used in pattern analysis and related areas such as data classification and data recognition before, and recognized its role for the importance of establishing a much-needed theoretical framework for visualization.

2.1. State-of-the-art of Pattern Analysis and data classification and recognition in marine data

We started our efforts with a carefully selected list of important publications in interactive machine learning and marine data. Using these candidate documents, we first tried an open approach to coding to identify "interesting" features at the intersection of computer vision and machine learning in general. Although this resulted in a high-level structure [19], it was impractical to make the analysis more concrete. Therefore, we decided to analyze a much larger set of sample articles, with two implications for our methodological options. (1) We understood the need to look at specific pattern analysis problem (in our scenario, intelligent driving, image synthesis, and object pose measurement) searching to make the research more focused, practical, and concrete. (2) We needed automated methods to narrow down the pool of potentially interesting articles.

During this process, we repaired the retrieval practice in cleaning its criteria and coding selection multiple times. Our final workflow consisted of four main steps, shown in Figure 2: 1.) Obtaining ontology research trends, 2.) Reviewing the previous research and application of pattern analysis, 3.) Identifying marine data classification and recognition issues, 4.) Searching for pattern analysis and machine learning parameters to encode for a large part of the ontologies' semantic content.

2.2. Sample network of records

Our common goal was the ontology research developed and how its implementation interacts in the pattern analysis and marine data communities. We decided to take a representative sample of papers, made up of every paper ever published in a Web of Science (WoS) source titles in the marine pattern analysis community from 1991 to 2021. From the database (WoS), is defined a collection of pattern analysis (PA) and marine data records (6048), data classification (DC) and marine data papers (3242) and data recognition (DR) and marine data records (1214), for a total of 9,899 records.

2.2.1. Paper metadata-based filtering

Methodological options were driven by the idea that the state of the art of ontology (machine/deep learning) research could be determined by using metadata. By metadata, we refer to aspects of the words-in-title that were deemed essential to facilitate a meaningful analysis in a full-content context. The initial synthesis was accomplished by deciding on a uniform list of metadata and their distribution along the years, as found in Figure 1. Based on this metadata definition, we implemented metadata lists from the sets of records in PA and data classification and recognition. The final metadata lists and statistics from this metadata filtering process are provided in Figures 2a,b,c.

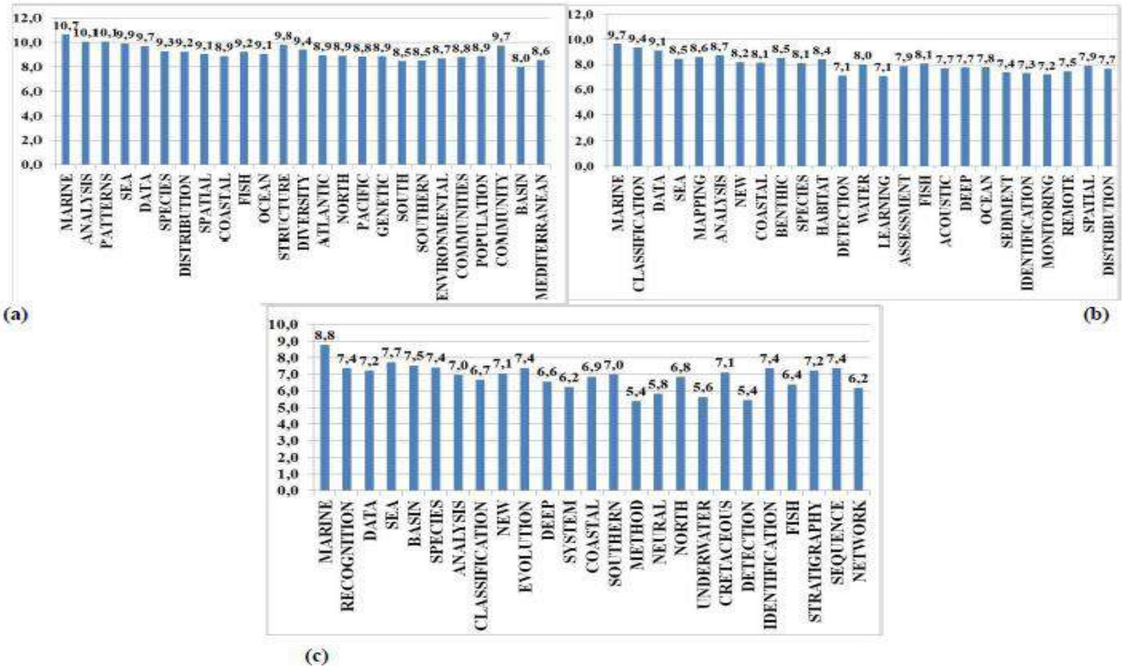


Figure 2: Comparison of metadata efforts in tracing WoS records to measure pattern analysis, marine data classification and marine data recognition shown in three log-scale histograms; 25 top metadata required with pattern analysis capture the data to detect, recognize and identify target of interests from physical, optical, fluid, and chemical underwater parameters (a); histogram estimated by image or video parameters with an emphasis on multilevel data classification is reported by its 25 top metadata (b); ranking title terms (metadata) of the documents on data recognition (c)

We formed a set of primary papers in marine ontologies with and without the initial criteria on PA and machine intelligence for data classification and recognition. Not all of these metadata allow to express the semantic content of an image. The discrepancy between the visual presentation of the image and its meaning requires machine learning performance expressed in terms of semantic resources as a ontology. So from the data set, including records in PA and corresponding DC and DR values (9899) are extracted two test sets on ontologies (42) and machine learning (210). In this way, the use of marine ontologies as the data classification and recognition technique focuses on the viability of using ontologies to solve the problem of pattern analysis.

2.2.2. Manual and automatic sample check

The 42 ontological papers were manually checked using the following criteria. First, we checked if the paper is a theoretical and evaluative framework or if it deals with a combination of applied or technical visual methods; as we planned to build a theoretical model for visualization. Second, we checked whether the paper addresses the combination of pattern analysis (PA), data classification (DC) and data recognition (DR), and whether the interaction returns to the visualization area. This had the advantage to present an interesting one workflow for the multi-source, multi-format, multi-dimension characteristic of marine data. Moreover, there is a return to the visualization area in its framework design that considers underlying data patterns. Given our focus on visualization we include this model that even feedback to the analysis of the 3D marine data. One major advantage of this method is its ability to define a semantic model of the issue under scrutinize (PA, DC & DR) combined with the associated domain of visualization to list the data visualization theories brought by marine data and observations, that range from the digital transformation of operational oceanography processes to the adoption of artificial intelligence. On this basis, we manually analyzed the first 42 candidate relevant documents obtained. Table 2 provides a partial list of the 42 specific ontological contexts detected in the PA and DC and DR data sources, and the extent to which they provide the ontology tools they use.

Table 2

List of the ontological contexts detected in the PA and DC and DR data sources (A/B: applied//theoretical; DV: data visualization feedback)

Num	Experimental Organism	A/B	DV	Ontology
1	Dynamena pumila	B	Y	Gene Ontology (GO) and KEGG pathway
2	Takifugu rubripes	A	Y	Gene Ontology (GO)
3	phytoplankton	A	Y	Gene Ontology (GO)
4	nd	A	Y	nd
5	Dreissena polymorpha	B	Y	Gene Ontology (GO)
6	Atlantic salmon	A	Y	Gene Ontology (GO) and UniProt Knowledgebase
7	Micromonas polaris; Pyramimonas tychotreta	B	Y	Gene Ontology (GO)
8	Crassostrea gigas	A	Y	Gene Ontology (GO)
9	Nd	B	Y	Genomic Standards Consortium's MIxS and Environment Ontology (ENVO); EMP Ontology (EMPO) of microbial environments
10	Chlamys farreri	A	Y	Gene Ontology (GO) and Eukaryotic Orthologous Groups (KOG) and Kyoto Encyclopedia of Genes and Genomes (KEGG)
11	Nd	B	Y	Protégé environment (ontology)
12	Nd	B	Y	Protégé environment (ontology)
13	Eucheuma denticulatum	A	Y	Gene ontology (GO)
14	48 species of freshwater and marine fish	A	Y	Gene ontology (GO)
15	Larimichthys crocea	A	Y	Gene ontology (GO)
16	Seriola lalandi	B	Y	Gene ontology (GO)
17	marine and FW sticklebacks	B	Y	Gene ontology (GO)
18	Genypterus chilensis	A	Y	Gene ontology (GO)
19	Ceriops	A	Y	Gene ontology (GO)
20	Human	A	Y	Gene ontology (GO)
21	Zostera muelleri	B	N	Gene ontology (GO)
22	Nd	B	Y	Integrated Ocean Observatory System ontology and the Marine Metadata Interoperability ontology (MMI)
23	Nd	A	Y	SWEET (Semantic Web for Earth and Environmental Terminology)
24	Mytillus galloprovincialis; Crassostrea-gigas; Chlamys farreri	A	Y	Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG)
25	Human	B	Y	BRENDA Tissue Ontology (BTO); Gene ontology (GO); human anatomy atlas CAVEman
26	Tachypleus tridentatus	A	Y	Gene ontology (GO)
27	Nd	B	Y	Uberon metazoananatomy ontology; Environmental Ontology (EnvO)
28	Orcinus orca	B	N	Gene ontology (GO)
29	Perna viridis; Mytilus galloprovincialis; Manila clam	B	Y	Gene ontology (GO)
30	Nd	B	N	nd
31	Pinctada martensii	A	N	Gene ontology (GO)
32	Mytilus galloprovincialis	A	N	Gene ontology (GO)
33	Sea cucumbers	B	Y	Gene ontology (GO)
34	Epinephelus coioides	A	Y	Gene ontology (GO)
35	Sparus aurata	B	Y	Gene ontology (GO)
36	Drosophila	B	Y	Gene ontology (GO); Kyoto Encyclopedia of Genes and Genomes (KEGG)
37	Mytius galloprovincialis	A	Y	Gene ontology (GO)
38	mouse brain	A	Y	Gene ontology (GO); KEGG pathway; web tool DAVID
39	Pacific oyster Crassostrea gigas	A	Y	GENE ontology (GO); program KAAS (KEGG Automatic Annotation Server)
40	adult male mice	A	Y	Gene ontology (GO); web tool DAVID (The Database for Annotation, Visualization and Integrated Discovery
41	dogfish shark (Squalus. acanthias)	A	Y	Gene ontology (GO); PANTHER gene ontology classification system (Applied Biosystems)
42	Nd	B	N	nd

After an automated process, the set of 210 papers corresponding to machine learning was filtered based on the fact that one of the most frequently used data visualization techniques in machine learning is the histogram plot. ML-based data visualization techniques were approached through metadata generated from a base histogram and classified into four levels: disseminative, observational, analytical and model-developmental. That is to say, a theoretical framework, because a visualization technique that builds on machine learning therefore attests its power for interactive analysis of heterogeneous marine data, it can deliver relevant pattern analysis content in the appropriate mode. Table 3 lists these visualization levels. Through boxplot with the ontological and machine learning (ML) datasets, we found differential expression of how their values are spread out and detect schematically their outliers (Figure 3). The temporal analysis parameters for machine learning (ML) are listed in Table 4.

Table 3
Levels of data visualization in machine learning methods for pattern analysis (marine data) (metadata generated from a base histogram

	analytics	association	case study	classification	comparison	complex	correlation	development	information	mapping	model	monitor	observation	pattern	prediction	recognition	regression	sensor	simulation	time series	other	SUM
Papers	1	1	4	70	7	3	1	4	2	16	20	12	2	13	13	7	1	12	3	2	16	210
disseminative				1					2			4		1						2	3	13
observational			1	11								2	2	1				12			4	33
analytical	1	1	3	34	7	3	1					6		10			1				3	70
model-developmental				24				4		16	20			1	13	7			3		4	92

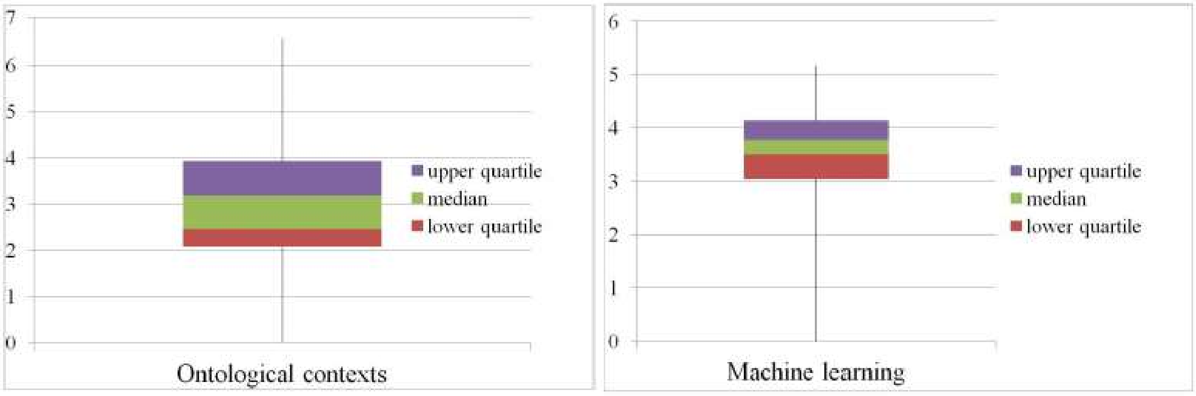


Figure 3: Differential expressions of relevance from the Ontological and Machine learning (ML) datasets (dependent variable: ln(cit))

Table 4
Machine learning for pattern analysis time statistics (marine data)

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	SUM
Papers	3	3	7	5	4	6	9	12	16	15	27	55	48	210
disseminative							1		1	1	1	5	4	13
observational	1		1			1		1			6	12	11	33
Analytical	1	1	3	2	1	4	4	6	6	6	7	14	17	70
model-developmental	1	2	3	3	3	1	4	5	9	8	13	24	16	92

3. Results and discussion

3.1. Ontology research trends review

The outputs of review using the ontology research trends are shown in Figure 1. As mentioned in Section 2.1, database entities represented as ontology terms result in a rich variety of scenarios that store in annotations their features and strengths. The ultimate goal is a system that combines visual and textual semantics to regularly annotate video sequences final aim is a system that will link the visual and text semantics in order to routinely annotate video sequences with the appropriate keywords of a domain expert. Most ontology-based cognitive vision promising results occurred in 2008.

By 2015, heterogeneous marine data, believed by visualization techniques to be of strategic importance, had their top priority. It is also the subject of scientific research, as evidenced by a large number of research papers, books, and reports. Highlights were the initiative to create in marine systems well-founded ontologies embedding these domain semantic and logical frameworks in the underwater environments thus providing opportunities for intelligent observatory units. The details of communication ontology that can be used by Remotely Operated Vehicles (ROVs) to transmit data and commands between vehicles and operators is defined by OWL in SWARMs platform. SWARMs users can estimate the rate of spread of pollutants and determine the level of pollution and the estimated size of the polluted area. In marine biology, search engines do use ontologies like SWEET for engaging the coral reef research community via a cyberinfrastructure network.

3.2. State of the art of pattern analysis in marine data classification and recognition

The outputs selection was initially based on the idea that machine learning (ML) enhanced by the ontology is able to compare pattern analysis performances using marine data. But ontologies specific coverage statistics are few, and it is difficult to say what actually constitutes a significant part of the terms in an ontology.

Typically, a generic ontology design pattern is developed for data from observations on the Semantic Web by unlocking the potential of compositional definitions, proposed to distinguish reliable relations for pattern analysis expression. These definitions are the necessary information to start with, because they are partitioned into mutually exclusive cross-products sets, many of which reference other candidate ontologies for chemical entities, proteins, biological qualities and zoological entities. An example of such a case is the Environment Ontology (ENVO) which, using the expressivity of OWL, grows in acceptance and participation in new user communities, thus offering an example of ontology's classes increased granularity in their logical definitions, allowing more flexibility in semantically advanced questions, inferences and analysis.

Ontologies such as the Extensive Observation Ontology (OBOE), Observations and Measurements (O&M), Semantic Sensor Network (SSN), and SWEET can be interconnected and expanded to include additional concepts more specific to the field of remote sensing, including the basic concepts that remote sensing professionals rely on to interpret remote sensing images (e.g., concepts, associated with spectral bands, spectra or texture indices). Examples of such a remote sensing ontology have already been applied, but have not yet been used in any upper ontology.

In this sense, under a definition of data classification as a process of clustering these data into a series of groups or categories, regardless of the method used for this purpose, research on pattern analysis finds a framework for marine data based on ontologies as an active output for computer vision. From this framework for an ontology-based ocean image classification that describes how to create ontological models for low- and high-level features, classifiers and rule-based expert systems, a much larger set of sources appeared.

3.3. Histograms to bridge the semantic gap between notions of content and similarity

The results from the previous section are used to build a sample of filtered dataset (9,899 records). Firstly, such sample contains metadata from thousands of word-in-title terms that can capture both concrete and abstract relationships between salient visual properties. Subsequently, histogram analysis methods were employed to compare the semantic effort by considering the metadata weight as generated on the base of global citation scores. This result in three referential frames, is shown in Figure 2a,b,c.

Structure (292 records) and communities (199) have a significant score in pattern analysis, indicating that these two parameters cause a positive effect on modelling required to discriminate relevant from non-relevant images (Figure 2a). Basin has a weak score due to the indirect value it has with the general purpose retrieval of features constructs such as predicates, relations, conjunctions, and a specification syntax for image content (for instance, photographic images).

The following is the histogram for data classification (see Figure 2b), which shows that mapping (221 records) is the main technique to classify marine data, while spatial (93) data are by now less rigidly circumscribed. The complete histogram of the attributes obtained for quantify the features in the data recognition domain is also shown Figure 2c. The importance of data recognition for evolutionary biologists (53 records) is enclosed within the scope of the study of species (67 records). The close relation with classification (61 records) ensures that a visual language can use an important mechanism for conscious control, limiting the range of possible configurations of functions that must be taken into account when performing a visual recognition task.

3.3.1. A framework for interactive visual analysis of heterogeneous marine data

The statistical quantized histogram metadata analysis was based on the available PA, DC & DR data, and focused on expressing the multi-dimension spatiotemporal marine data in one workflow. Based on the data we processed, two visualization methods are explored: ontology and machine learning (ML). The basic idea is shown in Fig. 3.

In this way, according to the data value and methodological choice, two different data classification and recognition scenarios in pattern analysis can be compared. Under the first scenario ontology-based for image retrieval and annotation was used to derive marine data patterns. Owing to the substantial positive bias in ontological feedback to the domain of visualization (Y=85.7%; Table 2), subsequent approaches for visual events were larger than in the case of the second scenario (machine learning (ML)), but because citations were restricted to 70% of the available ML data, the resulting lower quartile of 0.45 reached a best score for ML than for ontologies (0.38).

Therefore, machine learning (ML) decisions based on marine data assessment outperformed ontology-driven coding for image classification. And that, in spite of ontology mapping for underwater IoT (IoUT) supports better interoperability protocols in the context of computer vision.

3.3.2. PA, DC & DR identified through ontologies for marine data

The Table 2 is built based on the 42 selected ontology records. This is a general-purpose scheme designed for filtering the typology of the data sources. As described by the PubMed database there is a strong proximity between applied (or technical) and theory (evaluative, comparative, lessons) contents when they are both expressed in percentage terms (47.61% vs. 52.35). This analysis is then extended utilizing the feedback of each source to the domain of data visualization. It is overly positive (85%), as estimated by subject headings including in MeSH for each source.

For all the sources that have been used in this study, the ontology tools are listed in Table 2. Researchers proposed new models to cope with marine transcriptome/genome identification (80%; Table 2). They assumed Gene ontology (GO) approaches to model knowledge on the experimentation organisms. Some approaches focus on a data visualization package (Illumina sequencing technology) to provide refined descriptions of the whole scenario (1,5,18,19,24,36; Table 2). In one source (39; Table 2), the authors propose an automatic reasoning mechanism to deal with uncertainty in a quasi-empirical model using KAAS automatic annotation server. A number of marine or environmental ontologies (MMI, ENVO, EMPO, SWEET) are found (9,11,12; Table 2), they are used as dictionary learning in microbial environments to find out unavailable variables. Other ontologies' tools are also used (BRENDA Tissue Ontology (BTO)(25; Table 2), Protein Annotation through Evolutionary Relationships (PANTHER) (17,41; Table 2).

3.3.3. Machine learning levels of visualization and their temporal perspective

In this section, it is investigated whether a data visualization level can give a prediction of its suitability for a particular machine learning task. There exists a spectrum of different steps of visualization ranging from high abstraction levels (e.g., model-developmental tools) to lower levels (e.g., operational aids) (see Table 3; 210 records). To enhance this theoretical framework performance, the ML-based data visualization techniques are used based on 20 metadata, assuming that the marine data papers are categorized from different histograms which are quite reliable.

On Table 3, the level of visualization importance for suitability prediction is shown. As the first basic task of knowledge discovery, it proposes the use of data classification tools (33%); most visual analytics processes reported in PA, DC & DR literature operate at this level. The analyst knows or assumes the model to be correct only in 6% of the sources. Only in 16% human analysts need to use visualization to observe data routinely. Human analysts are able to observe input data in conjunction with the machine's "understanding" in many ways (33%). A line-up of model developmental tools gains a new understanding in terms of complexity (44%).

We can further derive from Table 3 that deep learning is the main automated support of marine data analytics using machine learning (ML) techniques, perhaps this is caused by the success of deep learning in computer vision tasks (eg. image classification, object detection, instance segmentation). In data visualization most of the deep learning studies focus on model developmental aids (44%), followed by observational tools (30%), investigative (18%) and presentational (8%) aids. Analytical and model-developmental visualization levels were shared equally among other ML techniques employed (transfer learning, ensemble methods, clustering).

As shown in Table 4, a sequence has been used to encode the data sources as the number of papers published by year. Years 2020 (55) and 2021 (48) are peaks. We can see clearly that 50% of the papers were all published in the last two years. This is not strange since the idea that in machine learning (ML) pattern analysis is gaining future, is expressed again by the importance of the two last years for the four levels of visualization (relative importance of 69%, 70%, 44%, 43%). As expected from the results in Table 3, there are gaps in both disseminative and observational data sets (with 6 years breaks in between).

4. Conclusions

This proof of principle study explores the potential uses of ontologies to encode for marine data pattern analysis literature. This study focuses on characterizing marine ontologies to select data visualization techniques. The underlying assumption is that the application of ontologies in marine data poses the problem of how should marine databases be represented. Therefore, the validation against pattern analysis in oceanography should be first put in terms of data interoperability. Using this approach could provide experts with a tool and method where they can rate ocean technologies and how they have been received in the communities where they have been placed. A data histogram approach has been adopted, which draws on the analysis of literature until significant categories appear. This has demonstrated its worth in pattern analysis, data classification and data recognition, and is regarded as an ingredient of the new generation of theoretical frameworks for data visualization. The results of the model to predict what the encoding for a large part of the ontologies' semantic content is going to look like in the future show that marine ontologies specific coverage statistics are few. It is acknowledged that the biomedical Gene Ontology (GO) currently represents the most successful implementation of ontologies in the domain of oceanography for pattern analysis applications including data visualization. It is recognized that, for machine learning data visualization, marine data scoring solutions were better than ontology-based coding for image classification. This approach has led to accurate predictions of the level of visualization importance for the example of data classification. Over the machine learning techniques most used for computer vision tasks with marine data, the result of the study outstands for it is clearly stated that deep learning is a promissory approach to gain new understandings in terms of data visualization tools. The results of this study show the potential use of marine data for pattern analysis assessment and prediction of the level of data visualization. This method shows the potential of ontologies to support the generation of model scenarios for image retrieval and annotation, and to aid for the empirical assessment of machine learning techniques. A single example data visualization was used as an application for indicating the potential value of ontologies to solve the issues of pattern analysis and taking a first step towards a theoretical model for visualization with marine data. It is recommended for future research that marine model developers should intensify their efforts to improve data discovery, sharing, and integration on the base of ontologies.

5. References

[1] K. Malde, N.O. Handegard, L. Eikvil, A.B. Salberg, Machine intelligence and the data-driven future of marine science, ICES J. Mar. Sci. 77 (2020) 1274-1285. doi:10.1093/icesjms/fsz057.

[2] T.R. Gruber, A translation approach to portable ontology specifications, Knowl.Acquis. 5 (1993) 199-220. doi:10.1006/knac.1993.1008.

[3] A. Ferreira, S. Marini, M. Attene, M.J. Fonseca, M. Spagnuolo, J.A. Jorge, B. Falcidieno, Thesaurus-based 3D object retrieval with part-in-whole matching, Int. J. Comput. Vis. 89 (2010) 327-347. doi: 10.1007/s11263-009-0257-6.

[4] M. Zurowietz, T.W. Nattkemper, Unsupervised knowledge transfer for object detection in marine environmental monitoring and exploration, IEEE Access 4 (2020). doi: 10.1109/ACCESS.2020.3014441

[5] B. Kamgar-Parsi, J.L. Jones, A. Rosenfeld, Registration of multiple overlapping range images - scenes without distinctive features, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 857-871. doi: 10.1109/34.93805.

[6] K. Ramar, T.T. Mirnalinee, An ontological representation for Tsunami early warning system, in: Proc. Int. Conf. Adv. Eng., Sci. Manage. (ICAESM), Nagapattinam, Tamil Nadu, India, 30–31 March 2012; IEEE, 2012, pp. 93-98.

[7] R. Lou, Z. Lv, S. Dang, T. Su, X. Li, Application of machine learning in ocean data. Multimed. Syst. (2021). doi:10.1007/s00530-020-00733-x.

[8] N. Boucquey, K.St. Martin, L. Fairbanks, L.M. Campbell, S. Wise, Ocean data portals: performing a new infrastructure for ocean governance, Environ. Plann. D 37 (2019) 484-503. doi: 10.1177/0263775818822829

A complete list of references is available from the author.

Concrete Names for Complex Expressions in Ontologies: A Survey of Biomedical Ontologies

Christian Kindermann[1], Martin Georg Skjæveland[2]

[1]Stanford University, 450 Serra Mall, Stanford, USA

[2]University of Oslo, Problemveien 7, 0315 Oslo, Norway

Abstract

The representation of an entity in an ontology may require complex expressions to capture all of its relevant characteristics. If an entity can be *defined* based on its characteristics, then its definition can be explicitly stated in most knowledge representation languages, such as the Web Ontology Language (OWL). Specifically, a domain-specific entity can be identified by a name in an ontology, which can be declared to be logically equivalent to a more complex expression This not only fixes the meaning of the entity in the ontology but also allows its name to replace the more complex expression throughout the ontology. Consistently using concise and informative names for domain-specific entities in an ontology can arguably enhance ontology comprehension, maintenance, and usability in practice. This raises the question of the extent to which entities represented in ontologies are associated with concrete names and how such names are used.

In this paper, we analyze how often named classes in OWL ontologies are defined as logically equivalent to complex expressions. We investigate whether such named classes are consistently used whenever possible and whether they are associated with labels intended for human understanding. Our findings indicate that complex class expressions are frequently declared to be equivalent to named classes in ontologies, and that such named classes are linked to human readable labels. While there seems to be a tendency to encourage the reuse of these names, we also observe a notable number of instances where such named classes are not consistently reused despite being defined.

Keywords

Ontology Engineering, Biomedical Ontology, Web Ontology Language, OWL

1. Introduction

The representation of an entity in an ontology typically involves statements about the entity's characteristics. When an entity can be defined based on its characteristics, the definition may include an informative name by which the entity can be referred to. Specifically, an entity's name may be used instead of its more complex definitional description. Despite the potential benefits of consistently using concise and informative names whenever possible, it has been observed that this practice is not always followed in published ontologies. To illustrate this, we revisit a concrete example taken from the Galen ontology, which was originally presented by Nikitina and Koopmann [1]. Here, the medical concept Clotting is represented as follows:

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: christian.kindermann@stanford.edu (C. Kindermann); martige@ifi.uio.no (M. G. Skjæveland)

ORCID: 0000-0001-7818-3466 (C. Kindermann); 0000-0002-9736-8316 (M. G. Skjæveland)

$$\begin{aligned}\mathsf{Clotting} \equiv\ & \exists\ \mathsf{actsSpecificallyOn}.(\mathsf{Blood}\\ & \quad \sqcap \exists\ \mathsf{hasPhysicalState}.(\mathsf{PhysicalState} \sqcap \exists\ \mathsf{hasState}.\mathsf{Liquid}))\\ & \sqcap \exists\ \mathsf{hasOutcome}.\mathsf{SolidBlood}\end{aligned}$$

This axiom is arguably complex due to both its size and the nesting of expressions. However, Galen also contains the following axioms:

$$\begin{aligned}\mathsf{LiquidBlood} &\equiv \mathsf{Blood} \sqcap \exists\ \mathsf{hasPhysicalState}.\mathsf{LiquidState}\\ \mathsf{LiquidState} &\equiv \mathsf{PhysicalState} \sqcap \exists\ \mathsf{hasState}.\mathsf{Liquid}\end{aligned}$$

Given these equivalences, the named concept LiquidBlood can be used to simplify the representation of Clotting to

$$\begin{aligned}\mathsf{Clotting} \equiv\ & \exists\ \mathsf{actsSpecificallyOn}.\mathsf{LiquidBlood}\\ & \sqcap \exists\ \mathsf{hasOutcome}.\mathsf{SolidBlood}\end{aligned}$$

The latter representation of Clotting is arguably easier to read, comprehend, and maintain. This observation raises questions about *the frequency of defining* concrete names for complex expressions, *the consistency of using* such names throughout an ontology, and to what extent the use of names *simplifies* the definition of more complex concepts. The contributions presented in this paper are as follows: (i) we propose an approach for identifying named classes with logical definitions in ontologies, (ii) we develop techniques for quantifying the use and lack of reuse of such named classes, and (iii) we use these techniques to conduct an empirical investigation on a large and complex corpus of ontologies in the biomedical domain to shed light on the use of such names in real-world ontologies.

2. Preliminaries

We assume the reader to be familiar with OWL [2] and only fix some terminology. Let N_C, N_I, and N_P be sets of *class names*, *individual names*, and *property names*. A *class* is either a class name or a *complex class* built using OWL class constructors. We will use $\top$ and $\bot$ to denote `owl:Thing` and `owl:Nothing` respectively. We use both OWL Functional Style Syntax [3] and Manchester Syntax [4] to write OWL axioms. An ontology is a set of axioms and we write $\mathcal{O} \models \alpha$ to denote that the ontology $\mathcal{O}$ entails the axiom α. An axiom α is *explicit* in $\mathcal{O}$ if $\alpha \in \mathcal{O}$, and *implicit* if $\alpha \notin \mathcal{O}$ but $\mathcal{O} \models \alpha$. An OWL expression e *occurs* in $\mathcal{O}$ if e is used as a subexpression within an explicit axiom in $\mathcal{O}$.

3. Abbreviations in Ontologies

The Oxford English Dictionary defines the word *abbreviation* to denote "[t]he result of shortening something; an abbreviated or condensed form, esp. of a text; a summary, an abridgement" [5]. So, we define an abbreviation for a complex OWL expression in terms of an equivalent named class. More formally, let A be a named class and C be a complex class expression. Then A is an

$\alpha_1 =$ SpicyPizza EquivalentTo Pizza and
(hasTopping some (PizzaTopping and (hasSpiciness some Hot)))
$\alpha_2 =$ SpicyTopping EquivalentTo PizzaTopping and (hasSpiciness some Hot)
$\alpha_3 =$ SpicyTopping EquivalentTo HotTopping
$\alpha_4 =$ DiavolaPizza SubClassOf SpicyPizza
$\alpha_5 =$ DiavolaPizza SubClassOf Pizza and hasCountryOfOrigin value Italy
$\alpha_6 =$ NapoletanaPizza SubClassOf Pizza and hasCountryOfOrigin value Italy

Figure 1: Example of abbreviations and synonyms in a sample ontology. The named class SpicyPizza is an abbreviation. The classes SpicyTopping and HotTopping are synonyms.

abbreviation for C in an ontology $\mathcal{O}$, if $\mathcal{O} \models \mathit{EquivalentClasses}(\mathsf{A}, \mathsf{C})$. We will refer to the $\mathit{EquivalentClasses}$ axiom as the **definition** of the abbreviation A.

A complex OWL expression can be equivalent to more than just one named class. We refer to equivalent named classes as *synonyms*.[1] In particular, a **synonym** for a named class N in an ontology $\mathcal{O}$ is a named class S s.t. $\mathcal{O} \models \mathit{EquivalentClasses}(\mathsf{S}, \mathsf{N})$ and we will refer to the $\mathit{EquivalentClasses}$ axiom as the synonym's definition. Please note that synonyms are not necessarily abbreviations. However, a synonym for an abbreviation is also an abbreviation (due to transitivity of $\mathit{EquivalentClasses}$).

Both abbreviations and synonyms are notions based on entailment, i.e., an *EquivalentClasses* axiom with exactly two arguments. However, OWL specifies *EquivalentClasses* as an n-ary constructor. So, for the purpose of analyzing how abbreviations and synonyms are specified in ontologies, we introduce the notion **syntactic definitions types** for both abbreviations and synonyms. In particular, an axiom of the form $\mathit{EquivalentClasses}(\mathsf{A}, \mathsf{C})$ and $\mathit{EquivalentClasses}(\mathsf{S}, \mathsf{A})$ will be referred to as **simple definitions** for the abbreviation A and the synonym S respectively. An axiom of the form $\mathit{EquivalentClasses}(\mathsf{A}, \mathsf{C}_1, \dots, \mathsf{C}_n)$ where $\mathsf{C}_1, \dots \mathsf{C}_n$ are complex class expressions is a **ambiguous definition** of A. An axiom of the form $\mathit{EquivalentClasses}(\mathsf{S}_1, \dots, \mathsf{S}_m)$ is an **enumerative definition** for the synonyms $\mathsf{S}_1, \dots, \mathsf{S}_n$. And lastly, an axiom of the form $\mathit{EquivalentClasses}(\mathsf{S}_1, \dots, \mathsf{S}_m, \mathsf{C}_1, \dots, \mathsf{C}_n)$ will be referred to as a **compound definition** for $\mathsf{S}_1, \dots, \mathsf{S}_m$, which are both synonyms and abbreviations.

With this notion of definition types, we can quantify how abbreviations and synonyms are specified explicitly in an ontology. However, counting implicit definitions of abbreviations and synonyms is not as straightforward, as extracting finite sets of entailments is a non-trivial matter [7]. We will delve into the determination and counting of implicit abbreviations and synonyms in more detail in Section 4. Before that, though, we address the more obvious question of how abbreviations and synonyms are *used* in an ontology.

Consider the example ontology $\mathcal{O}_{Ex}$ shown in Figure 1. Here, the abbreviation SpicyPizza is specified via a simple definition in axiom α_1 and occurs on the right-hand side of α_4. So, we say an abbreviation is **used** if it occurs in an OWL axiom that is not its definition. In addition to the *use* of an abbreviation, we can also determine if an abbreviation is *not used* even though its use would be possible. We refer to such a case as an abbreviation's **possible use**. For example,

[1] The Oxford English Dictionary defines the word *synonym* to denote "Strictly, a word having the same sense as another (in the same language); [...]" [6].

consider axiom $\alpha_1 \in \mathcal{O}_{Ex}$. Here, the abbreviation $\mathsf{SpicyTopping}$ (and its synonym $\mathsf{HotTopping}$) has *possible uses* since the complex OWL expression $\mathsf{PizzaTopping}$ and ($\mathsf{hasSpiciness}$ some Hot) could be replaced by either $\mathsf{SpicyTopping}$ or $\mathsf{HotTopping}$.

With the notions of an abbreviation's use and possible use, we can quantify the impact of abbreviations in an ontology. Before we do so, we come back to the topic of *determining* both explicit and implicit definitions of abbreviation and synonyms in an ontology.

4. Determining Abbreviations and Synonyms

Explicit definitions for abbreviations can be easily determined by checking the syntactic shape of all axioms in a given ontology. Similarly, implicit definitions can be determined by checking $\mathcal{O} \models \mathit{EquivalentClasses}(\mathsf{A}, \mathsf{C})$ for all pairs of named classes and complex classes occurring in an ontology. However, this becomes impractical for large ontologies with numerous named and complex classes.

Instead, to determine implicit abbreviations, we build upon highly optimized implementations of the standard reasoning service *classification*, i.e., computing all entailed $\mathit{SubClassOf}$ and $\mathit{EquivalentClasses}$ axioms between named classes in an ontology [8, 9, 10]. We will refer to this set as the inferred class hierarchy (ICH). The idea is to introduce an abbreviation for every complex class expressions that occurs in a given ontology, then to compute the ICH of the ontology with these newly added abbreviations, and finally to read off all implicit abbreviations from the ICH.

More formally, for a given ontology $\mathcal{O}$, we create the *abbreviation ontology*

$$\mathcal{O}_A = \mathcal{O} \cup \{\mathit{EquivalentClasses}(\mathsf{A}_i, \mathsf{C}_i) \mid \mathsf{C}_i \text{ occurs in } \mathcal{O}, \mathsf{A}_i \text{ does not occur in } \mathcal{O}\}$$

and compute $\mathsf{ICH}(\mathcal{O}_A)$. Since the ICH captures all $\mathit{SubClassOf}$ and $\mathit{EquivalentClasses}$ relationships between named classes in an ontology, it is straightforward to identify all named classes in $\mathcal{O}$ that are equivalent to a newly introduced abbreviation A_i in $\mathcal{O}_A$.

We will demonstrate this procedure by way of example. Consider the ontology $\mathcal{O}_{Ex}$ shown in Figure 1. This ontology contains complex class expressions $\mathsf{C}_1, \ldots, \mathsf{C}_6$ as shown in Figure 2a. Classifying the abbreviation ontology $\mathcal{O}_{Ex}^A$ and inspecting the ICH (see Figure 2) reveals that, for example, $\mathsf{SpicyTopping}$ is equivalent to A_2, which in turn is equivalent to C_2 by construction. So, $\mathsf{SpicyTopping}$ is an abbreviation for C_2 in $\mathcal{O}_{Ex}$.

5. Study Design and Materials

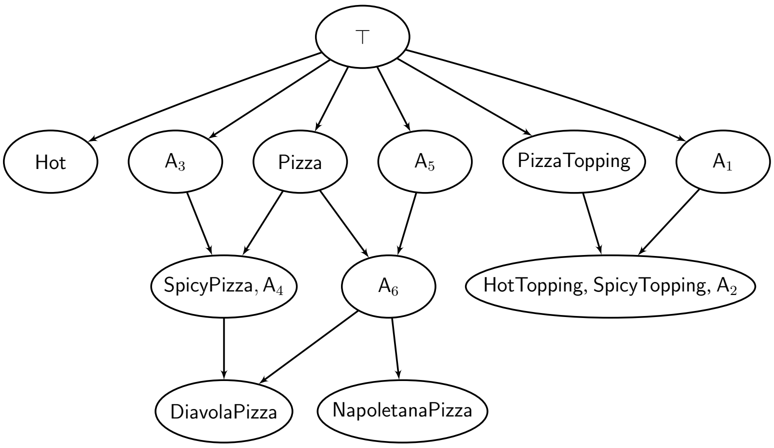
Before we investigate to what extent abbreviations and synonyms are defined, used, and not used even though this would be possible, we first establish a baseline. This baseline aims to determine whether named classes in ontologies are associated with concrete domain-specific terms that are intended for human interpretation. Specifically, we assess the association of named classes in ontologies with human-readable annotations specified via `rdfs:label`[2] and `obo:definition`.[3] Similarly, we establish a baseline for human-readable synonyms specified

[2] https://www.w3.org/TR/rdf-schema/#ch_label

[3] http://purl.obolibrary.org/obo/IAO_0000115

$C_1 =$ hasSpiciness some Hot
$C_2 =$ PizzaTopping and (hasSpiciness some Hot)
$C_3 =$ hasTopping some (PizzaTopping and (hasSpiciness some Hot))
$C_4 =$ Pizza and (hasTopping some (PizzaTopping and (hasSpiciness some Hot)))
$C_5 =$ hasCountryOfOrigin value Italy
$C_6 =$ Pizza and hasCountryOfOrigin value Italy

(a) Complex class expressions in $\mathcal{O}_{Ex}$.



(b) Visualisation of ICH($\mathcal{O}^{A}_{Ex}$) without $\bot$.

Figure 2: Determining implicit abbreviations in $\mathcal{O}_{Ex}$ via the inferred class hierarchy of $\mathcal{O}^{A}_{Ex}$.

via `skos:altLabel`,[4] `oio:hasExactSynonym`, `oio:hasNarrowSynonym`, `oio:hasBroadSynonym`, `oio:hasRelatedSynonym`,[5] or `obo:alternativeLabel`.[6]

We conduct our empirical investigation using ontologies indexed in BioPortal as of February 2023.[7] The data set is created folllowing the same approach as described by Matentzoglu and Parsia [11] and includes a total of 785 ontologies. For orchestrating the empirical investigation, we use use the OWL API (v.5.1.15). We exclude ontologies that cannot be processed with the OWL API. Additionally, we exclude ontologies that do not contain any class expression axioms since such ontologies cannot contain abbreviations or synonyms. As a result of this procedure, our study corpus consists of 744 ontologies.

We group ontologies into three disjoint categories. First, ontologies that consist of atomic axioms only, i.e., *SubClassOf* and *EquivalentClasses* axioms that have only named classes as arguments. Second, ontologies expressible in $\mathcal{EL}^{++}$, and third, ontologies not expressible in $\mathcal{EL}^{++}$. We refer to these three kinds of ontologies as atomic, $\mathcal{EL}^{++}$, and rich ontologies

[4] https://www.w3.org/2012/09/odrl/semantic/draft/doco/skos_altLabel.html

[5] https://raw.githubusercontent.com/geneontology/go-ontology/master/contrib/oboInOwl#{hasExactSynonym, hasNarrowSynonym, hasBroadSynonym, hasRelatedSynonym}.

[6] http://purl.obolibrary.org/obo/IAO_0000118

[7] https://bioportal.bioontology.org/

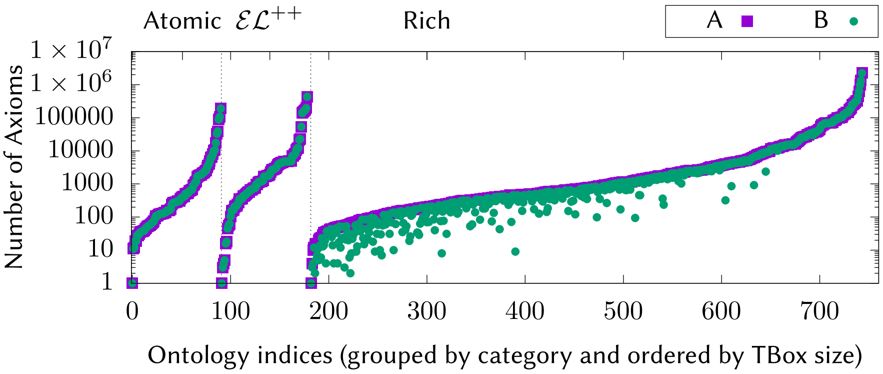


Figure 3: Size of (A) an ontology's TBox and (B) the set of an ontology's class expression axioms.

respectively. The study corpus contains 91 atomic ontologies, 88 $\mathcal{EL}^{++}$ ontologies, and 565 rich ontologies. We order ontologies within a category by the size of their TBoxes and assign each ontology an index in ascending order starting with atomic ontologies, then $\mathcal{EL}^{++}$ ontologies, and finally rich ontologies. Figure 3 illustrates this indexing by showing a comparison between the size of an ontology's (a) TBox and (b) the subset of class expression axioms.

Using the reasoner Konclude (v0.7.0-1138), we successfully classified 714 ontologies (for the purpose of determining implicit synonyms) and 656 abbreviation ontologies (for the purpose of determining implicit abbreviations).

6. Results

Before presenting our results on abbreviations and synonyms (see Section 3) we report on the use of annotation properties for specifying human-readable labels, definitions, and synonyms. Table 1 illustrates the number of ontologies that offer human-readable annotations for varying percentages of named classes.

We find that `rdfs:labels` are available in many ontologies for large proportions of named classes. For instance, $51+51+232 = 334$ ontologies provide `rdfs:labels` for all named classes. An additional $19 + 16 + 130 = 165$ ontologies provide `rdfs:labels` for at least 90% of named classes (but not 100%), so that $(334+165)/744 \approx 67\%$ of ontologies provide `rdfs:labels` for at least 90% of named classes. This provides strong evidence of the importance of human-readable `rdfs:labels` for named classes representing domain-specific concepts in biomedical ontologies.

We also find that `obo:definitions` are used in many ontologies. For example, 226 ontologies, i.e., $226/744 \approx 30\%$, provide `obo:definitions` for at least half of all named classes. While these proportions are smaller compared to `rdfs:labels`, they are non-trivial and provide evidence that `obo:definitions` play an important role in many biomedical ontologies.

However, human-readable annotations for synonyms appear to be less common in biomedical ontologies compared to `rdfs:labels` and `obo:definitions`. While there are a few ontologies

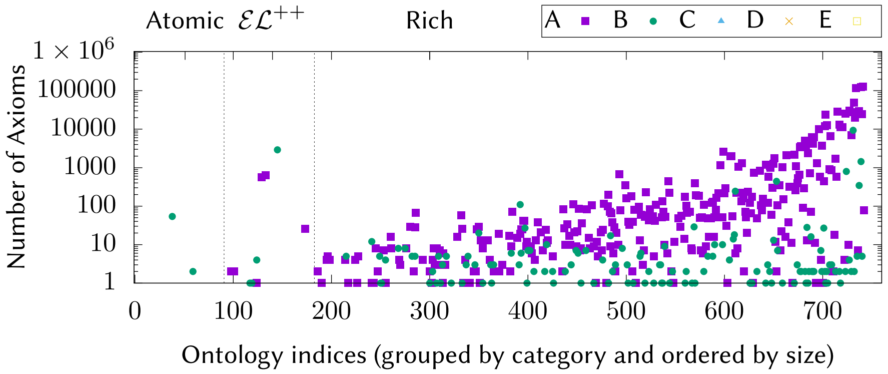


Figure 4: Number of (A) simple abbreviation definitions, (B) simple synonym definitions, (C) ambiguous abbreviation definitions, (D) enumerative synonym definitions, and (E) compound definitions.

Table 1
Number of atomic (A), $\mathcal{EL}^{++}$(E), and rich (R) ontologies that provide human-readable annotations for different proportions of named classes.

%	Number of Ontologies																							
	rdfs:label			obo:def			obo:alt			skos:alt			oio:exact			oio:narrow			oio:broad			oio:related		
	A	E	R	A	E	R	A	E	R	A	E	R	A	E	R	A	E	R	A	E	R	A	E	R
100	51	51	232	4	9	11	-	-	-	1	1	-	2	-	1	-	-	-	-	-	-	-	-	-
≥ 90	19	16	130	5	22	57	-	-	1	5	1	4	-	-	1	-	-	-	-	-	1	-	-	1
≥ 50	4	5	61	4	8	106	-	-	5	1	1	7	3	3	30	-	-	1	-	-	-	2	4	6
≥ 20	-	2	19	3	2	31	-	-	23	-	5	13	5	7	67	-	-	5	-	-	1	2	12	42
> 0	1	1	30	4	9	41	1	1	96	4	-	27	10	27	69	5	9	96	4	10	87	12	21	103
0	16	13	93	71	38	319	90	87	440	80	80	514	71	51	397	86	79	463	87	78	476	75	51	413

that annotate more than 90% of named classes with synonyms, e.g., 12 in the case of `skos:alt`, a lot of ontologies do not provide such synonym annotations for named classes at all (see last row in Table 1). This suggests that although annotations for synonyms are used in some biomedical ontologies, they do not seem to hold the same level of importance as `rdfs:labels` and `obo:definitions` for the most part.

The last observation can also be made w.r.t. the logical notions of abbreviations and synonyms. Figure 4 shows how many $\mathit{EquivalentClassesAxioms}$ are syntactic definition types for abbreviations or synonyms (see Section 3 for definition types). It becomes evident that there are about twice as many ontologies in which abbreviations are (explicitly) defined compared with ontologies in which synonyms are (explicitly) defined — namely 309 and 136 respectively. We also note that abbreviations and synonyms are specified only via simple definitions.

The difference between abbreviations and synonyms is not only evident in the *number of ontologies* in which they are defined but also in the *number of definitions* within ontologies. We find that the definitions for abbreviations are more numerous compared to definitions for synonyms. Specifically, there are 105 ontologies that define more than a hundred abbreviations, whereas only eight ontologies have more than a hundred definitions for synonyms. Given these observations, we will focus on abbreviations rather than synonyms in the remainder of this paper and will start with a discussion of explicitly defined abbreviations and then proceed with implicitly defined ones.

Figure 5 shows how many abbreviations are defined in ontologies. We observe that explicitly defined abbreviations (represented by green dots in Subfigure 5a) can be found in $309/744 \approx 41\%$ of ontologies. Furthermore, a considerable number of these ontologies contain numerous abbreviations, with 48 of them having at least a thousand explicitly defined abbreviations. Additionally, we notice that each explicitly defined abbreviation tends to be used (as shown by the blue triangles, indicating the number of used abbreviations, on top of the green dots, indicating the number of defined abbreviations in Subfigure 5a). Specifically, in 248 out of the 309 ontologies with explicitly defined abbreviations, all abbreviations are also used.

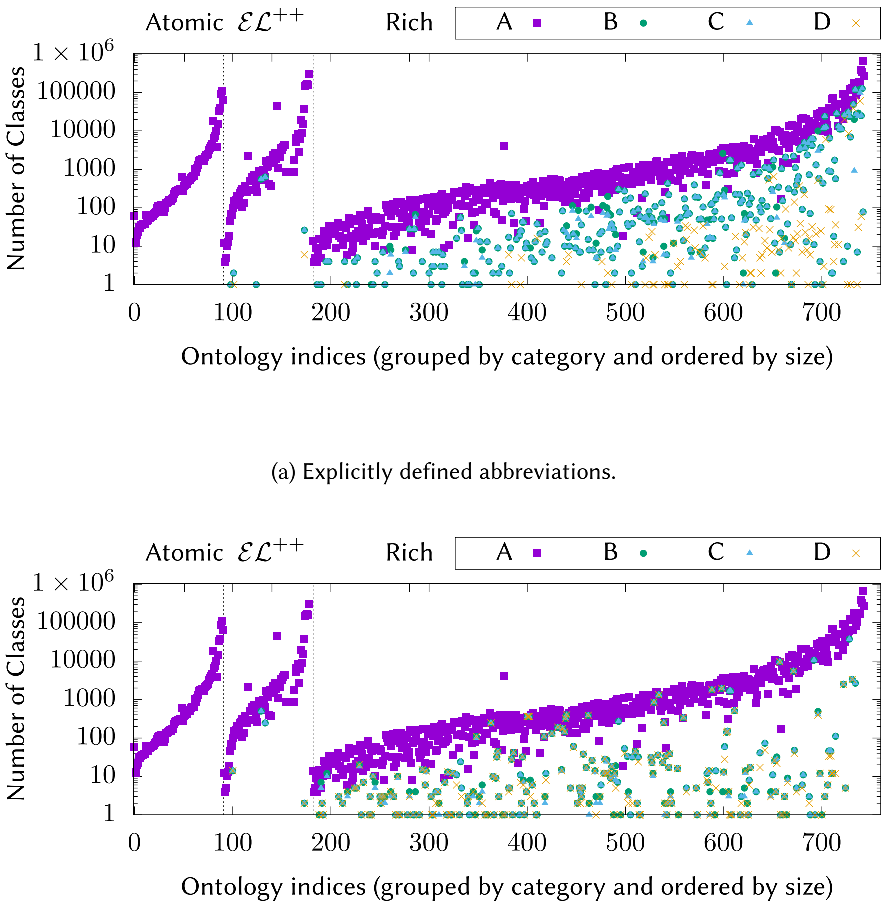
On the contrary, the number of explicitly defined abbreviations with potential uses tends to be considerably smaller compared to the overall number of defined abbreviations (indicated by the yellow cross in Subfigure 5a). Only one-third, specifically 101 out of 309 ontologies $(101/309 \approx 33\%)$, contain explicitly defined abbreviations with possible uses. Additionally, it is important to note that no ontology exists where *all* explicitly defined abbreviations have potential uses. These observations suggest that explicitly defined abbreviations are generally used whenever possible, but there are a few exceptions in which over a thousand explicitly defined abbreviations have potential uses.

Regarding implicit abbreviations, there are 231 ontologies that define at least one abbreviation implicitly. It seems that ontologies generally have fewer implicitly defined abbreviations compared to explicitly defined ones. For instance, there are only 36 ontologies with more than a hundred implicitly defined abbreviations. However, it is important to note that implicit abbreviations could not be computed for 88 ontologies, and this group include many larger ontologies.

Before presenting the number of uses and possible uses for implicitly defined abbreviations, we remind the reader of the definition of an abbreviation's use in an ontology. An abbreviation's use is considered as its occurrence outside of its definition. Now, an implicitly defined abbreviation necessarily occurs in the ontology but there is no explicit definition. So, any implicitly defined abbreviations is also used (the only exceptions being `owl:Thing` and `owl:Nothing`). Regarding possible uses, we find that almost all implicitly defined abbreviations come with possible uses. In other words, most implicitly defined abbreviations are not used in at least one case where it would be possible to use them.

Besides counting *how many* abbreviations are defined, used, or not used, we are also interested in the question of *how often* abbreviations are used or could possibly be used. Since reporting these numbers for all abbreviations defined in all ontologies would be impractical (considering that some ontologies contain several thousand abbreviations), we focus on reporting the data for each ontology regarding the abbreviations with the *most uses* and *most possible uses.* Figure 6 depicts these numbers for both explicit and implicit abbreviations. We find that the abbreviation with the highest use for both explicitly and implicitly defined abbreviations (represented with a purple square and blue triangle respectively in Figure 6) tends to fall between 10 and 100 in most ontologies. However, in some large ontologies, this number can be much higher, reaching several thousand.

In the case of most possible uses, we find that the numbers for implicit abbreviations (represented with a yellow cross in Figure 6) are larger compared to the numbers of explicit ones (represented with a green dot). This provides additional evidence that explicit abbreviations tend to be used where possible, whereas implicitly abbreviations are not consistently reused.



(a) Explicitly defined abbreviations.

(b) Implicitly defined abbreviations.

Figure 5: Number of (A) named classes, (B) classes defined as abbreviations, (C) used abbreviations, (D) abbreviations with possible uses.

7. Related Work

Logical equivalent rewritings for ontologies are usually motivated for the purpose of improved reasoning performance [12] or ontology-based data access [13]. However, the idea of rewriting axioms to improve ontology comprehension has also been discussed. Existing work in this direction focuses on rewritings that are minimal in size because large expressions are arguably hard to read and comprehend [14, 15, 1]. Yet, it is debatable whether the smallest possible logical rewriting of an axiom is indeed most suitable for human interpretation.

In the work presented in this paper, the focus is not on rewritings that are minimal in size. Rather, we study to what extent domain-specific vocabulary defined in an ontology can be *reused* to simplify otherwise complex expressions. The main argument being that a meaningful

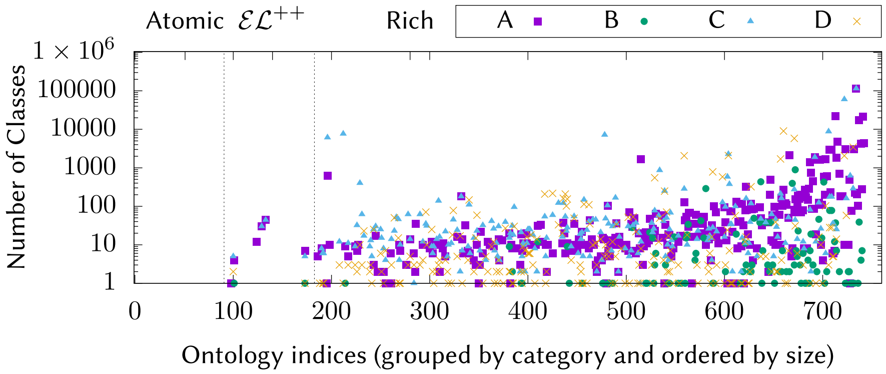


Figure 6: Number of maximal (A) explicit abbreviation use, (B) explicit abbreviation possible use, (C) implicit abbreviation use, (D) implicit abbreviation possible use.

name is more readily understood by domain experts compared to more complex expressions in OWL. It is important to note that the associated reduction in size is secondary in this context.

The task of determining abbreviations in an ontology (cf. Section 4) can be interpreted as concept definability, i.e., the problem of finding a definition for a concept name in an ontology [16]. However, we restrict the problem to finding definitions for concepts in terms of complex class expressions that already occur, syntactically speaking, in an ontology. Nevertheless, advances in research on concept definability may provide useful insights, e.g., knowing under what conditions implicitly defined concepts can also be defined explicitly.

8. Discussion & Future Work

The OBO Foundry considers naming conventions important for ontology comprehension, readability, navigability, alignment, and integration [17] and recommends that the majority of classes in an ontology should have textual definitions [18]. While not all biomedical ontologies conform to the principles put forward by the OBO Foundry, there is no question that human-readable names and definitions are used in many ontologies (see Table 1 in Section 6).

The use of human-readable names is important, because technical terms in a domain-specific vocabulary tend to be defined in terms of already defined terms. For example, a 'blood assay datum' is defined as 'A data item that is the specified output of a blood assay'. This definition only makes sense if the notion of a 'blood assay' is already defined. So, if textual definitions make use of already defined terms, and textual definitions should match logical definitions, as the OBO Foundry advocates, then possible uses of explicitly defined abbreviations (cf. Section 3) should not occur.[8]

However, our results suggest that even though the reuse of already defined concepts seems to be preferred, there is a non-trivial number of cases in which a complex class expression

[8] See https://obofoundry.org/principles/fp-006-textual-definitions.html.

could be replaced by an existing equivalent named class. It would be interesting to consult with the ontology developers in such cases to determine whether such cases are intended or not. Likewise, it would be interesting to find out whether classes with *implicit* logical definitions are intentional and should be made explicit, and whether they should be reused whenever possible.

In addition to the question of *when to use* an abbreviation, is the question of *when to introduce* a new abbreviation. In particular, if a complex class expression occurs often in an ontology, one may want to think about whether such an expression can be given a meaningful name and which should be used instead.

However, it needs to be highlighted that the introduction of an abbreviation, as defined in this work, *changes the meaning* of an ontology. Consider the ontology $\mathcal{O}$ and $\mathcal{O}_A = \mathcal{O} \cup \{\alpha\}$ where $\alpha = \mathit{EquivalentClasses}(\mathsf{A}, \mathsf{C})$ is a definition for an abbreviation A. If A does not occur in $\mathcal{O}$, then $\mathcal{O} \not\equiv \mathcal{O}_A$ because $\mathcal{O}_A \models \alpha$ but $\mathcal{O} \not\models \alpha$. This change in meaning can be avoided by encoding abbreviations using a meta-language, e.g., OTTR [19], on top of OWL. As an example, consider the ontology

$$\mathcal{O} = \{ \begin{array}{lllllll} \mathsf{Napoletana} & \mathsf{SubClassOf} & \mathsf{Pizza} & \mathsf{and} & \mathsf{hasCountryOfOrigin} & \mathsf{value} & \mathsf{Italy}, \\ \mathsf{Diavola} & \mathsf{SubClassOf} & \mathsf{Pizza} & \mathsf{and} & \mathsf{hasCountryOfOrigin} & \mathsf{value} & \mathsf{Italy}, \\ \mathsf{Hawaiian} & \mathsf{SubClassOf} & \mathsf{Pizza} & \mathsf{and} & \mathsf{hasCountryOfOrigin} & \mathsf{value} & \mathsf{Canada} \end{array} \}.$$

With OTTR, a mapping `ItalianPizza` $\mapsto$ Pizza and $\mathsf{hasCountryOfOrigin}$ value Italy can be defined, so that $\mathcal{O}$ can be encoded as

$$\mathcal{O}_T = \{ \begin{array}{lllllll} \mathsf{Napoletana} & \mathsf{SubClassOf} & \mathtt{ItalianPizza}, & & & & \\ \mathsf{Diavola} & \mathsf{SubClassOf} & \mathtt{ItalianPizza}, & & & & \\ \mathsf{Hawaiian} & \mathsf{SubClassOf} & \mathsf{Pizza} & \mathsf{and} & \mathsf{hasCountryOfOrigin} & \mathsf{value} & \mathsf{Canada} \end{array} \}.$$

Note that `ItalianPizza` is not an OWL class but an expression in OTTR. In particular, the ontology $\mathcal{O}$ is semantically equivalent to $\mathcal{O}_T$ because the OTTR expression `ItalianPizza` is *indistinguishable* from Pizza and $\mathsf{hasCountryOfOrigin}$ value Italy on the level of OWL. The use of a meta-level language also opens up possibilities to capture definitions on higher level of abstraction than OWL. In the case of the example ontology $\mathcal{O}$, the representation of a pizza's country of origin could be captured by a *parameterized* OTTR expression $\mathtt{PizzaWithOrigin}(x) \mapsto \mathsf{Pizza}$ and $\mathsf{hasCountryOfOrigin}$ value x. With this, all three pizzas in $\mathcal{O}$ can be encoded in a uniform manner giving rise to the following even more meaningful definitions:

$$\mathcal{O}_P = \{ \begin{array}{lll} \mathsf{Napoletana} & \mathsf{SubClassOf} & \mathtt{PizzaWithOrigin}(\mathsf{Italy}), \\ \mathsf{Diavola} & \mathsf{SubClassOf} & \mathtt{PizzaWithOrigin}(\mathsf{Italy}), \\ \mathsf{Hawaiian} & \mathsf{SubClassOf} & \mathtt{PizzaWithOrigin}(\mathsf{Canada}) \end{array} \}.$$

9. Conclusion

In this paper, we proposed an approach for analyzing and quantifying the use of logical abbreviations, i.e., named classes that are defined to be logically equivalent to complex class expressions. We used this approach to survey biomedical ontologies indexed in BioPortal and find that abbreviations are highly prevalent. Although there are some exceptions, *explicitly* defined abbreviations tend to be used whenever possible. However, *implicitly* defined abbreviations often come with many possible uses which rasies the question of whether this is intentional or undesireable.

References

[1] N. Nikitina, P. Koopmann, Small Is Beautiful: Computing Minimal Equivalent EL Concepts, in: AAAI, AAAI Press, 2017, pp. 1206–1212.

[2] B. Cuenca Grau, I. Horrocks, B. Motik, B. Parsia, P. F. Patel-Schneider, U. Sattler, OWL 2: The next step for OWL, Journal of Web Semantics 6 (2008) 309–322.

[3] B. Motik, P. Patel-Schneider, B. Parsia, OWL 2 Web Ontology Language. Structural Specification and Functional-Style Syntax (Second Edition) (2012). URL: http://www.w3.org/TR/owl2-syntax/.

[4] M. Horridge, P. Patel-Schneider, OWL 2 Web Ontology Language Manchester Syntax (Second Edition) (2012). URL: https://www.w3.org/TR/owl2-manchester-syntax/.

[5] abbreviation, n., in: OED Online, Oxford University Press, 2022. URL: https://www.oed.com/view/Entry/180?redirectedFrom=abbreviation&, accessed July 05, 2022.

[6] synonym, n., in: OED Online, Oxford University Press, 2022. URL: https://www.oed.com/view/Entry/196522?result=1&rskey=EVWsim&, accessed July 05, 2022.

[7] S. Bail, B. Parsia, U. Sattler, Extracting Finite Sets of Entailments from OWL Ontologies, in: Description Logics, volume 745 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2011.

[8] F. Baader, B. Hollunder, B. Nebel, H. Profitlich, E. Franconi, An Empirical Analysis of Optimization Techniques for Terminological Representation Systems, or Making KRIS Get a Move on, in: KR, Morgan Kaufmann, 1992, pp. 270–281.

[9] B. Glimm, I. Horrocks, B. Motik, R. D. C. Shearer, G. Stoilos, A novel approach to ontology classification, J. Web Semant. 14 (2012) 84–101.

[10] Y. Kazakov, M. Krötzsch, F. Simancik, The Incredible ELK - From Polynomial Procedures to Efficient Reasoning with $\mathcal{EL}$ Ontologies, J. Autom. Reason. 53 (2014) 1–61.

[11] N. Matentzoglu, B. Parsia, BioPortal Snapshot 30.03.2017, 2017. URL: https://doi.org/10.5281/zenodo.439510. doi:10.5281/zenodo.439510.

[12] D. Carral, C. Feier, B. C. Grau, P. Hitzler, I. Horrocks, *EL*-ifying Ontologies, in: IJCAR, volume 8562 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 464–479.

[13] M. Imprialou, G. Stoilos, B. C. Grau, Benchmarking Ontology-Based Query Rewriting Systems, in: AAAI, AAAI Press, 2012.

[14] F. Baader, R. Küsters, R. Molitor, Rewriting Concepts Using Terminologies, in: KR, Morgan Kaufmann, 2000, pp. 297–308.

[15] M. Horridge, B. Parsia, U. Sattler, Laconic and Precise Justifications in OWL, in: ISWC, volume 5318 of *Lecture Notes in Computer Science*, Springer, 2008, pp. 323–338.

[16] B. ten Cate, E. Franconi, I. Seylan, Beth Definability in Expressive Description Logics, J. Artif. Intell. Res. 48 (2013) 347–414.

[17] D. Schober, B. Smith, S. E. Lewis, W. Kusnierczyk, J. Lomax, C. Mungall, C. F. Taylor, P. Rocca-Serra, S. Sansone, Survey-based naming conventions for use in OBO Foundry ontology development, BMC Bioinform. 10 (2009). URL: https://doi.org/10.1186/1471-2105-10-125. doi:10.1186/1471-2105-10-125.

[18] R. C. Jackson, N. Matentzoglu, J. A. Overton, R. Vita, J. P. Balhoff, P. L. Buttigieg, S. Carbon, M. Courtot, A. D. Diehl, D. M. Dooley, W. D. Duncan, N. L. Harris, M. A. Haendel, S. E. Lewis, D. A. Natale, D. Osumi-Sutherland, A. Ruttenberg, L. M. Schriml, B. Smith, C. J. S. Jr., N. A. Vasilevsky, R. L. Walls, J. Zheng, C. J. Mungall, B. Peters, OBO Foundry in 2021: operationalizing open

data principles to evaluate ontologies, Database J. Biol. Databases Curation 2021 (2021). URL: https://doi.org/10.1093/database/baab069. doi:10.1093/database/baab069.

[19] M. G. Skjæveland, D. P. Lupp, L. H. Karlsen, H. Forssell, Practical Ontology Pattern Instantiation, Discovery, and Maintenance with Reasonable Ontology Templates, in: ISWC (1), volume 11136 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 477–494.

Leveraging Biomedical Ontologies to Boost Performance of BERT-Based Models for Answering Medical MCQs

Sahil Sahil[1,*], P Sreenivasa Kumar[1]

[1] *Department of Computer Science & Engineering, Indian Institute of Technology Madras, Chennai*

Abstract

Large-scale pretrained language models like BERT have shown promising results in various natural language processing tasks. However, these models do not benefit from the rich knowledge available in domain ontologies. In this work, we propose BioOntoBERT, a BERT-based model pretrained on multiple biomedical ontologies. We also introduce the Onto2Sen system to process various ontologies to generate lexical documents, such as entity names, synonyms and definitions, and concept relationship documents. We then incorporate these knowledge-rich documents during pretraining to enhance the model's "understanding" of the biomedical concepts. We evaluate our model on the MedMCQA dataset, a multiple-choice question-answering benchmark for the medical domain. Our experiments show that BioOntoBERT outperforms the baseline model BERT, SciBERT, BioBERT and PubMedBERT. BioOntoBERT achieves this performance improvement by incorporating only 158MB of ontology-generated data on top of the BERT model during pretraining, just 0.75% of data used in pretraining PubMedBERT. Our results demonstrate the effectiveness of incorporating biomedical ontologies in pretraining language models for the medical domain.

Keywords

Biomedical Ontologies, BERT, Medical Multiple Choice Question Answering

1. Introduction

Biomedical ontology research encompasses a variety of entities (from dictionaries of names for biological products to controlled vocabularies to principled knowledge structures) and processes (i.e., acquisition of ontological relations, integration of heterogeneous databases, use of ontologies for reasoning about biological knowledge) [1]. Biomedical ontologies include various aspects of medical terminologies such as symptoms, diagnosis and treatment.

Multiple-choice question-answering (MCQA) is a challenging task in general and in particular, in the domain of the medical field as the relevant knowledge is not commonly available in text corpora. The success of MCQA systems relies on striking a delicate balance between language understanding, domain-specific reasoning, and the incorporation of rich knowledge sources.

In the medical domain, the use of ontology-based QA systems has a very good potential to effectively capture domain-specific knowledge and provide accurate responses to medical

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*] Corresponding author.

cs20s017@cse.iitm.ac.in (S. Sahil); psk@cse.iitm.ac.in (P. S. Kumar)

0009-0004-0167-5621 (S. Sahil); 0000-0003-2283-7728 (P. S. Kumar)

queries. By harnessing biomedical ontologies, these systems can depict intricate relationships among medical concepts, resulting in more precise and contextually aware answers.

Ontology-based multiple-choice question-answering systems are few in number, but Ontology-based QA systems have shown promise in capturing domain-specific knowledge and accurately answering medical questions [2] [3]. By leveraging biomedical ontologies, these systems can represent complex relationships between medical concepts, enabling more precise and contextually aware responses. A major limitation is that using these systems requires an understanding of the ontology structure in order to formulate queries. For example, queries may necessitate using intermediate concepts in the ontology when there is no direct relationship between the concepts in question.

Contextual word embedding models, such as BERT (Bidirectional Encoder Representations from Transformers) [4] have achieved state-of-the-art results in many NLP tasks. Initially tested in a general domain, models such as BioBERT[5], UmlsBERT [6], SciBERT [7], and PubmedBERT [8], have also been successfully applied in the biomedical domain by pretraining them on biomedical corpora. However, current biomedical applications of transformer-based NLP models do not incorporate structured expert domain knowledge from a biomedical ontology into their embedding pretraining process.

To illustrate the significance of biomedical ontology knowledge, let's consider a scenario where a medical question pertains to a specific rare disease. While a pretrained language model trained on a vast corpus may have encountered related terms or phrases, it may lack the medical domain-specific knowledge required to provide accurate and nuanced answers. In contrast, a biomedical ontology encompasses structured and domain-specific knowledge, including relationships, hierarchies, and semantic information about medical concepts. By integrating such ontology knowledge into our models, we can tap into a comprehensive and precise representation of medical domain knowledge, enabling more accurate and contextualized question-answering.

In light of this research gap, our study aims to bridge the divide between ontology-based approaches and deep learning models in the context of MCQA in the medical domain. Specifically, our objectives are:

- To overcome the challenges of ontology injection, including the computational overhead and annotation burden associated with large biomedical ontologies.
- To investigate techniques for integrating biomedical ontological knowledge with pretrained BERT models in MCQA systems.

In this paper, we present a novel approach that bridges the gap between ontology-based methods and pretrained language models, harnessing the strengths of both to enhance multiple-choice question-answering (MCQA) in the medical domain. Our contributions to this work can be summarized as follows:

1. Onto2Sen, a simple yet effective solution for Ontology Injection: We propose a unique solution called Onto2Sen system to generate a comprehensive ontology-backed sentence corpus, which serves as a valuable resource for enriching pretrained models with domain-specific knowledge. By incorporating this rich semantic information from biomedical domain ontologies into the models, we anticipate enhancing their contextual understanding and reasoning abilities.

2. Introducing BioOntoBERT: We propose BioOntoBERT, a pretrained BERT model that leverages various Biomedical Ontologies using the Onto2Sen generated corpus. BioOntoBERT surpasses several other biomedical BERT models, including PubmedBERT [8], SciBERT[7] and BioBERT [5], in terms of performance for multiple-choice question answering on the MedMCQA dataset.

Furthermore, BioOntoBERT demonstrates remarkable performance with just 158MB of pretraining data, significantly reducing the computational cost and carbon footprint associated with larger models. This aspect makes our novel approach not only effective but also environmentally friendly, addressing the growing concerns regarding energy consumption in deep learning models and highlighting the power of knowledge.

2. Related Work

Biomedical Multiple Choice Question Answering (MCQA) is a significant task in natural language processing. Various approaches have been proposed to improve the performance of MCQA systems by leveraging ontologies and pretrained language models.

As mentioned earlier, Ontology-based MCQA models are relatively limited, while Ontology-based question-answering systems have shown promise in capturing domain-specific knowledge and providing accurate answers to medical questions. For instance, in the case of XMQAS proposed by Midhunlal et al.[9], the system utilized natural language processing techniques and ontology-based analysis to process medical queries and extract relevant information from medical documents. Other approaches, like the one presented by Kwon et al.[10] for stroke-related knowledge retrieval, employed SPARQL templates and medical knowledge QA query ontology to transform queries into executable SPARQL queries for retrieving medical knowledge. However, these approaches have limitations due to their reliance on a template-based approach, which may restrict the flexibility and adaptability of the system.

In addition to ontology-based approaches, using pretrained models has significantly advanced MCQA systems. One notable example is PubmedBERT [8], a variant of BERT designed explicitly for biomedical text comprehension. These pretrained models, including PubmedBERT, have showcased remarkable performance in capturing medical terminologies and comprehending complex medical questions. Moreover, models like BioBERT [5], SciBERT[7], and UmlsBERT [6] have been finetuned for biomedical NLP tasks, exhibiting improved performance in various medical question-answering and information retrieval tasks. It is worth noting that these models are pretrained on extensive corpora, such as Pubmed abstracts entire medical dataset, which consists of over 3.1 billion words.

Less amount of work has been done in using external knowledge with neural networks in the biomedical multiple choice question answering domain, whereas in other domains like common sense reasoning several different approaches have been investigated for leveraging external knowledge sources. Sap et al.[11] introduce the ATOMIC graph with 877k textual descriptions of inferential knowledge (e.g. if-then relation) to answer causal questions. Lv et al.[12] propose to extract evidence from both structured knowledge bases such as ConceptNet and Wikipedia text and conduct graph-based representation and inference for commonsense reasoning.

He et al.[13] proposed a training procedure to infuse disease knowledge and augment pretrained BERT models. Their experiments demonstrated improved performance in consumer health question answering, medical language inference, and disease name recognition. This motivates us to leverage the strengths of ontology which excel at representing complex medical concepts and terminologies. By integrating ontology and BERT-based models, we aim to enhance the capabilities of our MCQA system and improve its accuracy and effectiveness in addressing biomedical questions.

To bridge the gap between ontology-based approaches and deep learning models, the authors of [14] [15] [16] have explored techniques for ontology injection and infusing context. These approaches aim to enhance the models' language understanding and domain-specific reasoning capabilities by injecting ontological information into the models by modifying or adding new BERT layers or mapping the concepts and relationships of the ontology to the data. However, these models face various challenges in processing and incorporating large biomedical ontologies. The computational overhead required to handle and integrate the vast knowledge in such ontologies can be significantly high. Moreover, the process of mapping the ontology with the dataset and preparing annotated data demands substantial time and labour resources. The manual effort required for this task can be burdensome, hindering the scalability and practicality of these approaches.

3. Biomedical Ontologies

Biomedical ontologies play a critical role in the field of medicine by organizing and representing knowledge related to diseases, genes, anatomical structures, and medical concepts. They establish a standardized framework that captures and integrates information, promoting data sharing, interoperability, and knowledge discovery. We now briefly describe the prominent biomedical ontologies we use for our model:

1. **Disease Ontology (DO)** [17] (v1.2): The Disease Ontology is a standardized ontology created to offer the biomedical community consistent, reusable, and sustainable descriptions of human disease terms, phenotype characteristics, and related medical vocabulary disease concepts.
2. **Gene Ontology (GO)** [18] (v2023-04-01): It is a widely used ontology that focuses on representing the functional attributes of genes and gene products across different species. GO encompasses three main domains: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC). BP describes biological processes in which genes are involved, MF represents the molecular functions they perform, and CC defines their cellular locations.
3. **Foundational Model of Anatomy Ontology (FMAO)** [19] (v5.0.0): FMAO is an ontology that aims to represent human anatomy in a detailed and structured manner. FMAO provides a hierarchical organization of anatomical structures, capturing spatial relationships and functional associations between different body parts.
4. **Precision Medicine Ontology** [20] (v4.0): It is a comprehensive ontology that represents medical concepts and their relationships in a standardized manner. Medicine Ontology covers various medical domains, including diseases, symptoms, treatments, diagnostic procedures, and medical devices.

Table 1
Different Biomedical Ontologies used

Ontology	Scope	Classes	# Object Properties	# Annotations	# subClass
FMAO Ontology	Anatomy	104721	139	51	262548
Bioassay Ontology	Pharmacology	904	17	34	981
Dental Ontology	Dentistry	2745	62	28	6507
Gene Ontology	Bioinformatics	84108	297	60	192606
Precision Medicine Ontology	Medicine	76155	95	23	122760
Disease Ontology	Pathology	11033	2	53	11063
Paediatrics Ontology	Paediatrics	1771	-	8	1760
HPS Ontology	Physiology	2920	86	34	3143
Mental Disease Ontology	Psychiatry	879	41	102	940

5. **Bioassay Ontology (BAO)** [21] (v1.1): The BAO focuses on establishing common reference metadata terms and definitions required for describing relevant information of low-and high-throughput drug and probe screening assays and results.
6. **Dental Ontology** [22] (v2016-06-27): It captures dental-related concepts and relationships, providing a standardized vocabulary for representing dental conditions, procedures, materials, and anatomical structures. It facilitates the integration of dental data and knowledge, supporting research, education, and clinical practice in dentistry.
7. **Pediatrics Ontology** (v2.0): Ontology focuses on representing pediatric healthcare-related concepts and their relationships. It covers various aspects of pediatric medicine, including diseases, developmental milestones, treatments, and interventions.
8. **Human Physiology Simulation Ontology (HPSO)** [23] (v1.1.1): HPSO captures the concepts and relationships related to the simulation and modelling of human physiology. It provides a standardized framework for representing physiological processes, organ interactions, and computational models.
9. **Mental Disease Ontology (MDO)** [24] (v2020-04-26): MDO represents mental disorders and related concepts. It offers a standardized vocabulary for categorizing and annotating mental diseases, symptoms, treatments, and diagnostic criteria.

4. Methodology

In this section, we present our approach for pretraining and fine-tuning a BERT[4] model on biomedical ontologies for multiple-choice question answering on the MedMCQA dataset. Our approach involves several key steps: data preparation, pretraining on biomedical ontologies, and fine-tuning the MedMCQA dataset. The code implementation is publicly available on GitHub[1].

4.1. Datasets

4.1.1. Multiple Choice Questions Dataset

We use the MedMCQA dataset[25], which consists of 1,94,000 multiple-choice questions on around 2400 healthcare topics and 21 medical subjects from one of the toughest entrance exams conducted for medical graduates in India, i.e., AIIMS and NEET PG. The diversity of questions

[1] https://github.com/sahillihas/BioOntoBERT

Table 2
Sample MCQA question from MedMCQA dataset with the correct answer as (A)

Question: Dentigerous cyst is likely to cause which neoplasia?	
(A) Ameloblastoma	(B) Adenocarcinoma
(C) Fibrosarcoma	(D) All of the above

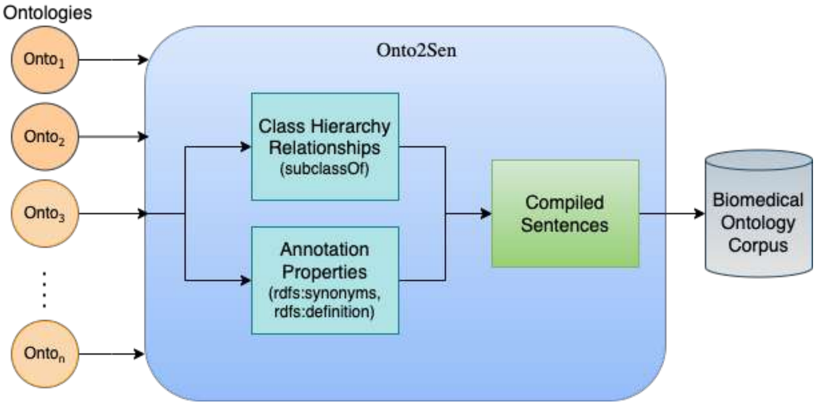

Ontologies
$Onto_1$
$Onto_2$
$Onto_3$
$Onto_n$
Onto2Sen
Class Hierarchy Relationships (subclassOf)
Annotation Properties (rdfs:synonyms, rdfs:definition)
Compiled Sentences
Biomedical Ontology Corpus

Figure 1: Proposed Onto2Sen Framework to generate BERT input corpus from the Ontologies

in the MedMCQA makes it a challenging dataset containing many aspects of medical knowledge; Table 2 illustrates one such example. Another distinguishing factor of this dataset is its questions are created for and by human experts. The dataset has three parts: the training set of 1,82,822 questions, the validation set of 4183 and the test set comprising 6150 questions, with an average token length of 12.35, 13.91 and 9.68, respectively. The answer choices are provided in the 'labels' column, encoded as integers 0, 1, 2, and 3. The ground truth for the test set is not publicly available. Hence we will be analysing the results on the validation set.

4.1.2. Ontology-based Sentence Generation

We propose a system called Onto2Sen to generate sentences from multiple ontologies curated from public resources mentioned in the previous section. It extracts concepts, annotations, and their properties from the ontology to form meaningful sentences. Onto2Sen preprocesses the ontologies and generates two types of sentences. The first type of sentence generated is from the subClass relationships. The second type of sentence is extracted from the relevant lexical annotation axioms in the ontology.
In the example shown in Figure 1, the Class Hierarchy Relationship sentences will contain the subClass property in the Disease Ontology (DO) allowing us to identify specific disease classifications. For instance, we can state that 'SPOAN syndrome is a neurodegenerative disease' using labels and identifiers in subClass relations. In addition, the transitive nature of the subclass properties is also utilized. Furthermore, Annotation Properties associated with diseases offer valuable insights into symptoms, synonyms and causal associations. For instance, we can describe that "SPOAN syndrome has synonym Spastic paraplegia" using the 'has_exact_synonym' annotation property.

We then used a natural language processing tool, spaCy, for preprocessing the compiled documents. We use these generated sentences as input to the model during pretraining to leverage the ontological knowledge.

After a study of the ontologies mentioned in Section 3, we find that using annotation properties and the class hierarchy for sentence generation is commonly applicable across all these ontologies and hence we adopt only these techniques for the present.

4.2. Pretraining Model

Pretraining is a crucial aspect of the BERT (Bidirectional Encoder Representations from Transformers) [4] model, which has revolutionized the field of natural language processing. In the context of BERT, pretraining refers to the initial phase where the model is trained on vast amounts of unlabeled text data, such as web documents or books. During this pretraining phase, BERT learns to generate contextualized representations of words and capture intricate semantic relationships by leveraging the bidirectional nature of transformers.

We propose a novel approach using Biomedical ontologies to pretrain the BERT model. As mentioned in the previous section, Onto2Sen can generate a corpus of meaningful sentences from different Biomedical ontologies. We use this generated corpus consisting of about 20M words which is a substantial volume of unlabeled text data related to the medical domain. The corpus was preprocessed and prepared for training, ensuring it was suitable for the subsequent steps.

The BERT model's pretraining phase involves two tasks: Masked Language Modelling (MLM) and Next Sentence Prediction. However, for our model, which incorporates biomedical ontologies, we focus on augmenting the Masked LM task and omit the Next Sentence Prediction task.

In the Masked LM task, we masked out 15 per cent of tokens in a sentence, and the model is trained to predict the original tokens given the context of the surrounding words. This approach will help the semantic understanding of medical terminologies by directly injecting biomedical ontology concepts and properties into the input sequence. As a result, the model will recognise and better understand medical concepts and terminologies effectively.

During the pretraining process, the BERT model was trained using the Adam optimizer, a widely adopted optimization algorithm for neural networks. The optimizer iteratively adjusted the model's parameters to minimize a predefined loss function, optimizing its ability to capture language patterns. Additionally, a learning rate scheduler was employed to dynamically adjust the learning rate at specific intervals, facilitating improved convergence and optimization of the model. The scheduler strategy, such as linear or exponential decay, was carefully selected based on experimentation and optimization.

These pretraining steps establish a well-built foundation for subsequent finetuning and proficient utilization of the BioOntoBERT model across diverse downstream natural language processing tasks.

4.3. Finetuning BERT

During the fine-tuning stage, we aim to train our BioOntoBERT model to accurately answer multiple-choice questions on the MedMCQA dataset without using any external context.

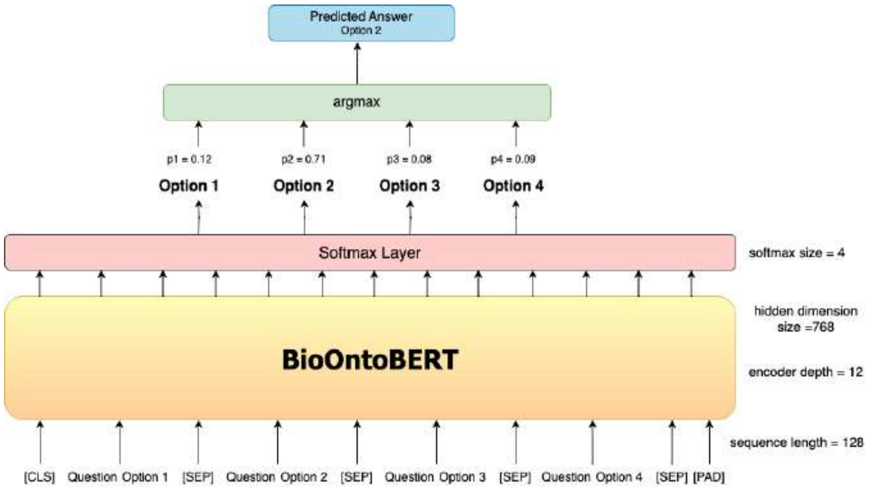


Figure 2: BioOntoBERT for multiple choice questions

Each multiple-choice question in the MedMCQA dataset was concatenated with its answer options to form a single input sequence of the form as shown in Figure 2.
Next, we performed tokenization on the dataset. Tokenization involves breaking down the questions and answers choices into smaller units called tokens, which the model can handle. This step ensures that the data is in a format suitable for the BioOntoBERT model to process. After the dataset is properly tokenized, we then train the BioOntoBERT model on this data.

During training, the model learned from the dataset by adjusting its internal parameters to better capture the relationships between questions and answer choices. The goal was to enhance the model's capacity to accurately choose the right answer when presented with a question. In this case, the labels were encoded in a one-hot format derived from integers. Throughout the training process, the model iteratively refined its understanding of the task by analyzing the patterns and context in the data. We carefully optimized the model's performance by adjusting various parameters, such as the learning rate and the number of training epochs.

Once the training was completed, we evaluated the performance of the finetuned BioOntoBERT model using the validation dataset. This evaluation allowed us to measure how well the model performed on unseen data and provided valuable insights into its ability to answer multiple-choice questions accurately.

During the fine-tuning process and subsequent evaluation of the BioOntoBERT model, a probability distribution is generated for each question's answer choices. The output probability distribution is denoted by $p1$, $p2$, $p3$ and $p4$ as shown in Figure 2. We identify the most likely answer choice by choosing the index associated with the highest probability.

5. Results

The main objective of this paper is to investigate the impact of incorporating biomedical ontology into the pretraining process of BERT models for the task of medical multiple-choice question answering. To achieve this objective, we developed a new pretrained model, BioOn-

Table 3
Accuracy and additional corpus size for different models on the MedMCQA dataset [25]. Statistics for prior BERT models are taken from their publications. [4] [5] [7] [8] .

Models	Corpus	Text Size	Accuracy
BERT	Wiki + Books	-	35%
BioBERT	PubMed	4.5B Words	38%
SciBERT	PMC + CS	3.2B words	39%
PubmedBERT	PubMed	3.1B words | 21GB	40%
BioOntoBERT (proposed)	Biomedical Ontologies	**20M words | 158 MB**	**42.72%**

toBERT, that is pretrained on a combination of 9 biomedical ontologies. We evaluated the performance of BioOntoBERT on the MedMCQA dataset, which contains a set of challenging medical questions curated by medical experts and compared it to the performance of other pretrained models, such as PubMedBERT[26], SciBERT[7] and BioBERT[5].

We conducted the pretraining of our BioOntoBERT model using the BERT base architecture, pretrained on English Wikipedia and BooksCorpus for 1M steps. BioOntoBERT was pretrained for 200K steps. The pretraining process involved a batch size of 32 and a learning rate scheduling of 5e-5. The pretraining and finetuning were both performed on a Tesla V100-PCIE-32GB GPU, with a maximum sequence length of 128. The pretraining of BioOntoBERT on ontology-generated sentences took approximately 10 hours only, whereas the pretraining times for PubmedBERT and BioBERT were reported as 5 days (120 hours) [8] and 10 days (240 hours) [5], respectively. For the finetuning process, a batch size of 32 and a learning rate of 1e-5 were selected. It took approximately 30 hours to complete the finetuning process due to the large size of the MedMCQA training data.

BioOntoBERT outperformed the baseline BERT-base, achieving a minimum accuracy of 42.72% in 10 runs. Furthermore, BioOntoBERT also outperformed PubMedBERT, which is pretrained on a huge corpus of biomedical text data. These results indicate that adding ontology data to the pretraining process can improve the performance of BERT models for medical question answering.

The comparison of models in Table 3 highlights the significance of the relatively small amount of additional ontology data we used to enhance the performance of our model. This finding suggests that the biomedical ontology we injected into the model is highly informative and beneficial, unlike much of the data in other corpora, which may be considered irrelevant.

During the evaluation, we also conducted a comparative analysis of the performance of BioOntoBERT, BERT, and PubmedBERT on various multiple-choice questions across different medical subjects. One evaluated question is in Table 2. Notably, BioOntoBERT correctly predicted the answer as (A) since the keywords 'Ameloblastoma', 'Adenocarcinoma', 'Fibrosarcoma' and 'Neoplasia' are present in the DOID ontology, BioOntoBERT model would have leveraged this knowledge. Whereas 'Dentigerous cyst' is not present in the DOID, Dentigerous cyst is a type of 'Odontogenic Cyst', and DOID contains a reference to 'Odontogenic Epithelium'. Odontogenic cysts and Odontogenic epithelium are closely related, as the former is derived from the remnants of the latter and forms as a result of abnormal developmental processes during tooth formation. In contrast, both BERT and PubmedBERT predicted the an-

Table 4
Subject-wise model comparison of PubMedBERT and BioOntoBERT on MedMCQA validation set of AIIMS MCQA. Statistics for PubMedBERT subject-wise are taken from [25]

Subject Name	PubMedBERT	BioOntoBERT	Ontology Used
Anatomy	39%	**41%**	✓
Biochemistry	49%	**50%**	✓
Dental	36%	**40%**	✓
ENT	**52%**	41%	✗
Medicine	47%	**48%**	✓
Microbiology	**44%**	40%	✗
Pathology	46%	**47%**	✓
Pharmacology	**46%**	42%	✓
Physiology	**56%**	54%	✓
Psychiatry	**56%**	50%	✓
Radiology	**31%**	28%	✗

swer as (D). This demonstrates an example instance of BioOntoBERT utilizing domain-specific knowledge.

The results presented in Table 4 demonstrate that BioOntoBERT exhibited superior performance compared to PubmedBERT across various subjects during pretraining, particularly when ontology data was available. Subjects like Anatomy, Biochemistry, Dental, Medicine, and Pathology showed notable improvements by including ontology data. However, for subjects such as ENT, Microbiology, and Radiology, where no ontology was used in our experiments, the benefits were not as evident. Additionally, for Pharmacology, Physiology and Psychiatry, the subject ontologies were not comprehensive enough to contribute significantly to question-answering capabilities. These findings underscore the significance of incorporating subject-specific ontology information to enhance the model's understanding and performance on domain-specific questions.

Importantly, we also evaluated the impact of the size and complexity of ontologies on the performance of the models. Surprisingly, we observed that the size or the number of concepts and properties in the ontologies did not necessarily correlate with improved question-answering performance. This suggests that the relevance and quality of the ontology data are crucial factors in enhancing the model's understanding and reasoning capabilities rather than the sheer quantity of information.

6. Conclusions

This study introduces the Onto2Sen system, which incorporates annotation-based and class-hierarchical sentences from ontologies to enhance the performance of a language model. It is the first instance of leveraging such knowledge in pretraining a language model for biomedical natural language processing tasks. The BioOntoBERT model, pretrained on biomedical ontologies, outperforms other models, including PubMedBERT, in multiple-choice question-answering tasks within the medical domain, effectively capturing medical terminologies. By achieving improved results with just 158MB of pretraining data, our approach not only enhances performance but also significantly reduces computational costs, making it a more sustainable approach to model training.

7. Future work

Firstly, the selection and incorporation of appropriate biomedical ontologies remain an ongoing challenge. While we employed several ontologies in our pretraining process, there are numerous other ontologies available that could potentially contribute to even better performance. Secondly, although BioOntoBERT exhibits impressive proficiency in language understanding and representation, it lacks advanced reasoning capabilities on ontologies. The model primarily captures contextual relationships and semantic information but does not possess explicit reasoning mechanisms to infer complex logical connections within ontologies. This limitation suggests avenues for future research, focusing on incorporating reasoning abilities into language models trained on biomedical ontologies.

References

[1] O. Bodenreider, A. Burgun, Biomedical ontologies, Medical Informatics: Knowledge Management and Data Mining in Biomedicine (2005) 211–236.

[2] Q. Guo, M. Zhang, Question answering based on pervasive agent ontology and semantic web, Knowledge-Based Systems 22 (2009) 443–448.

[3] A. Arbaaeen, A. Shah, Ontology-based approach to semantically enhanced question answering for closed domain: A review, Information 12 (2021) 200.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[5] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[6] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, A. Wong, Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus, arXiv preprint arXiv:2010.10391 (2020).

[7] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[9] M. Midhunlal, M. Gopika, Xmqas-an ontology based medical question answering system, International Journal of Advanced Research in Computer and Communication Engineering 5 (2016) 929–932.

[10] S. Kwon, J. Yu, S. Park, J.-A. Jun, C.-S. Pyo, Stroke medical ontology qa system for processing medical queries in natural language form, in: 2021 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2021, pp. 1649–1654.

[11] M. Sap, R. Le Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, Y. Choi, Atomic: An atlas of machine commonsense for if-then reasoning, in: Proceedings of the AAAI conference on artificial intelligence, volume 33, 2019, pp. 3027–3035.

[12] S. Lv, D. Guo, J. Xu, D. Tang, N. Duan, M. Gong, L. Shou, D. Jiang, G. Cao, S. Hu, Graph-

based reasoning over heterogeneous external knowledge for commonsense question answering, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 8449–8456.

[13] Y. He, Z. Zhu, Y. Zhang, Q. Chen, J. Caverlee, Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition, arXiv preprint arXiv:2010.03746 (2020).

[14] T. R. Goodwin, D. Demner-Fushman, Enhancing question answering by injecting ontological knowledge through regularization, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2020, NIH Public Access, 2020, p. 56.

[15] L. He, S. Zheng, T. Yang, F. Zhang, Klmo: Knowledge graph enhanced pretrained language model with fine-grained relationships, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4536–4542.

[16] K. Faldu, A. Sheth, P. Kikani, H. Akbari, Ki-bert: Infusing knowledge context for better language and domain understanding, arXiv preprint arXiv:2104.08145 (2021).

[17] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, W. A. Kibbe, Disease ontology: a backbone for disease semantic integration, Nucleic acids research 40 (2012) D940–D946.

[18] 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, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (2000) 25–29.

[19] C. Rosse, J. L. Mejino Jr, The foundational model of anatomy ontology, in: Anatomy ontologies for bioinformatics: principles and practice, Springer, 2008, pp. 59–117.

[20] L. Hou, M. Wu, H. Y. Kang, S. Zheng, L. Shen, Q. Qian, J. Li, Pmo: A knowledge representation model towards precision medicine, Math. Biosci. Eng 17 (2020) 4098–4114.

[21] U. Visser, S. Abeyruwan, U. Vempati, R. P. Smith, V. Lemmon, S. C. Schürer, Bioassay ontology (bao): a semantic description of bioassays and high-throughput screening results, BMC bioinformatics 12 (2011) 1–16.

[22] W. D. Duncan, T. Thyvalikakath, M. Haendel, C. Torniai, P. Hernandez, M. Song, A. Acharya, D. J. Caplan, T. Schleyer, A. Ruttenberg, Structuring, reuse and analysis of electronic dental data using the oral health and disease ontology, Journal of Biomedical Semantics 11 (2020) 1–19.

[23] M. Gündel, E. Younesi, A. Malhotra, J. Wang, H. Li, B. Zhang, B. de Bono, H.-T. Mevissen, M. Hofmann-Apitius, Hupson: the human physiology simulation ontology, Journal of biomedical semantics 4 (2013) 1–9.

[24] J. Hastings, W. Ceusters, M. Jensen, K. Mulligan, B. Smith, Representing mental functioning: Ontologies for mental health and disease (2012).

[25] A. Pal, L. K. Umapathi, M. Sankarasubbu, Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering, in: Conference on Health, Inference, and Learning, PMLR, 2022, pp. 248–260.

[26] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets, arXiv preprint arXiv:1906.05474 (2019).

The OntoWoH Ontology - a Women's Health Reference Ontology Detailing the Fragment of Climacteric and Menopause

Cilenir Carla de Carvalho[1], Andreia Soprani dos Santos[1], Silvia das Dores Rissino[1], Susana Bubach[1] and Maria das Graças da Silva Teixeira[1]

[1] *Federal University of Espírito Santo (UFES), BR-101, Km 60, Litorâneo, 29932-540, São Mateus, ES, Brazil*

Abstract

The reference ontology named Ontology for Women's Health (OntoWoH) is a representation of the complex and extensive domain of women's health, directed to climacteric and menopause aspects in its first version, based on the STRAW criteria. Women's health has repercussions in the cultural-social-legal-economic scenario, and it is the study object of several researches, with different approaches, due to its relevance. Considering the increase in women's life expectancy, the importance of these studies aimed at the phases of a woman's life from the climacteric onwards, extending to post-menopause, which need greater attention in public health systems, is highlighted. Thus, the goal of this work was to develop OntoWoH using the Systematic Approach for Building Ontologies (SABiO) methodology and based on the Unified Foundational Ontology (UFO) guidelines, with emphasis on ontology formalization activities and knowledge acquisition tasks. Among its contributions, this work shows the integration of Computer Science and Health Science areas, aiming to assist in the construction of a future quality information system in the field of Health Care, due to the increase of this type of information system in Brazil, mainly in the Unified Health System (SUS) (public healthcare assistance in Brazil), of which women are main users. In addition, OntoWoH was created to help understanding and communication in the stages of a woman's life, especially after the climacteric, with the intention of identifying a terminology (which may become standard) and paving the way for semantic interoperability in the domain. As next steps, OntoWoH should invest in reuse and integration activities, and can evolve into an operational ontology, initiating the mapping to a computational tool.

Keywords

UFO, SABiO, Reference Ontology, Women's Health, Climacteric, Menopause

1. Introduction

A conceptual model intends to improve understanding and communication of aspects in the physical and social world through a formal description [1]. Conceptual models bring numerous benefits to the representation of a domain and can be useful in several areas, such as Computer Science and the area of the domain (such as, Health Care or Law). One of the ways to build quality conceptual models is through Ontology-Driven Conceptual Modeling (ODCM).

On the other hand, the Health Care area applies Health Care Information Systems (HCIS) to obtain, manage and use information to improve health care tasks, increase the performance of services and facilitate their administration [2].

At same time, the demand for health information grows, and the challenges inherent to its use grow, such as knowledge sharing and data integration (including among different HCIS), which is essential for improving the quality of healthcare services [2, 3].

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: cilenir.carvalho@edu.ufes.br (C. C. Carvalho); andreia.s.santos@ufes.br (A. S. Santos); silvia.rissino@ufes.br (S. D. Rissino); susana.bubach@ufes.br (S. Bubach); maria.teixeira@ufes.br (M. G. S. Teixeira)

ORCID: 0009-0004-3129-2269 (C. C. Carvalho); 0000-0002-4377-6517 (A. S. Santos); 0000-0002-1180-9142 (S. D. Rissino); 0000-0002-7190-5275 (S. Bubach); 0000-0003-0398-8029 (M. G. S. Teixeira)

The construction of ontologies to represent a domain is one of the techniques to solve these problems [2]. The increase in the use of ontologies is justified by the fact that they provide a common conceptual framework that enables the development of shareable and reusable knowledge bases, which facilitates interoperability and the fusion of information, enabling the construction of powerful and more intelligent computational applications [4].

Once the benefits of ODCM have been established, it is necessary to identify the domain over which the model will be established. Keeping this in mind we chose to explore women's health.

Women have greater longevity than men, but their disability-free life expectancy has been lower, due to the increase in the occurrence of diseases such as depression, dementia and functional dependence (inability to maintain the physical and mental skills necessary for an independent and autonomous life), as well as cardiovascular [5, 6] and autoimmune [7] diseases. Considering the increase in woman´s life expectancy, there is a need for research and solutions focused on the phases of a woman's life, especially the climacteric and post-menopause, which also require greater attention in public health systems [8].

Climacteric and menopause occur in the lives of any woman, often triggering a series of symptoms that can cause suffering and misunderstanding of their families and other people, other than the affected woman herself. Despite advances in studies of Women's Health, there is still a significant lack of knowledge about the climacteric and menopause as a natural phenomenon in women's lives, which requires special care of health care professionals, with systematized and individualized guidance for women and close people [9].

Thus, this work aimed to understand the domain of Women's Health, focusing on the Climacteric and Menopause stage in its first version. Such understanding enabled the development of a consistent representation of the domain, which can be used by Health Care specialists, Computer Science specialists and other stakeholders, as an object of communication, deep understanding of the domain and, also, as income for Decision Support Information Systems, both for Women's Health specifically and for Health Care in general.

This paper is organized as follows. The first section contains the Introduction, which describes the context of the work, as well as the work itself. The second section contains the necessary theoretical foundation to understand the work and the main methodological aspects adopted in the development of OntoWoH. Then, section 3 presents the reference domain ontology. Concluding the paper, the last section describes our final considerations.

2. Theoretical Foundation and Methodological Aspects

2.1. The Domain of Women's Health - Climacteric and Menopause Stage

In Brazil, Women's Health was incorporated into National Health Care policies in the first decades of the 20th century, focusing on demands related to pregnancy and childbirth. Mother-infant programs, developed in the 30s, 50s and 70s, demonstrated a restricted view of women, taking into account their biological specificity and their social role as mothers and housewives, responsible for raising, educating and caring for their health of her children and other family members [10]. That is, demands for a comprehensive approach to women's health do not encompass all of their life stages, as strategies aimed at the female population occur mostly for women of reproductive age [11]. It is noteworthy that with the decline in birth rates and increased life expectancy, women in the climacteric phase may become the majority in health care assistance compared to pregnant women [12].

At present moment, there are still several gaps in relation to women's health care, such as: gynecological complaints; infertility and assisted reproduction; women's health in adolescence; chronic degenerative diseases; occupational health; mental health; infectious diseases and care during the climacteric/menopause [10].

Each phase of a woman's life has peculiarities. In addition to the phases of life common to any person (childhood, adolescence, adulthood and advanced age), it is observed that for women there may exist particularly feminine events: the menstrual cycle, pregnancy, the breastfeeding period and the climacteric stages (which includes menopause).

The terms menopause and climacteric, although used as synonyms, have different meanings [9]. According to the World Health Organization (WHO), climacteric is a biological phase of a woman's life, and not a pathological process, comprising the transition between a woman's reproductive and non-reproductive period. Menopause is a milestone of this phase and corresponds to the last menstrual cycle, being recognized after 12 months of its occurrence.

There is no predetermined age for menopause begins. Usually, it occurs around age 48 to 50 years [13], but it can start at 40 without being a problem [9]. When it occurs before the age of 45 (40 years for some authors), menopause is considered premature and can be caused by some chromosomal abnormalities, autoimmune disorders or even by some unknown cause. Also, it can be induced as a consequence of bilateral or iatrogenic surgical oophorectomy at any age. Late menopause occurs after the age of 55 years [14].

To better understand the menopause and climacteric stage, it is necessary to observe the previous phase of a woman's life. The menstrual cycle occurs during the female reproductive period, with two subcycles simultaneously - the ovarian, which occurs in the ovary, and the uterine, which occurs in the endometrium. The menstrual cycle occurs due to the function of the hypothalamic-pituitary-ovarian (HHO) axis. The hypothalamus stimulates the pituitary gland, which releases gonadotropins (follicle-stimulating hormone (FSH) and luteinizing hormone (LH)), which stimulate the ovaries. In contrast, hormones (estrogen and progesterone) produced by follicles in the ovaries regulate the hypothalamus, completing the cycle [15]. Menopause happens because, over time, there is a decrease in the number of follicles, a progressive drop in the concentrations of estrogen estradiol (E2) and progesterone, and the ovaries begin to respond less to the stimulus of the FSH and LH hormones. At first, women have more spaced menstrual cycles, then they stop ovulating and menstruation becomes irregular until they reach menopause [16].

Before the 2000s there was ambiguity in the use of the term pre-menopause by researchers: either it corresponded to the one or two years immediately preceding menopause; or comprising the entire reproductive period until the occurrence of menopause, which is the term recommended by WHO. Postmenopause corresponds to the period after menopause. Perimenopause corresponds to the period immediately prior to the occurrence of menopause, characterized by the onset of endocrine, biological and clinical changes that indicate the approach of menopause and extends to the first year after menopause; and the menopausal transition, which corresponds to the period of time before menopause, characterized by increased variation in the menstrual cycle. As the term climacteric was previously used interchangeably with perimenopause, to avoid confusion it was recommended that its use be dropped, but considering its popularity and dominance of the term climacteric, The Council of Affiliated Menopause Societies (CAMS) reinstated its use in 1999, determining it as a period that marks the transition from the reproductive to the non-reproductive period and that incorporates perimenopause, extending over a longer variable period, both before and after perimenopause [17, 18].

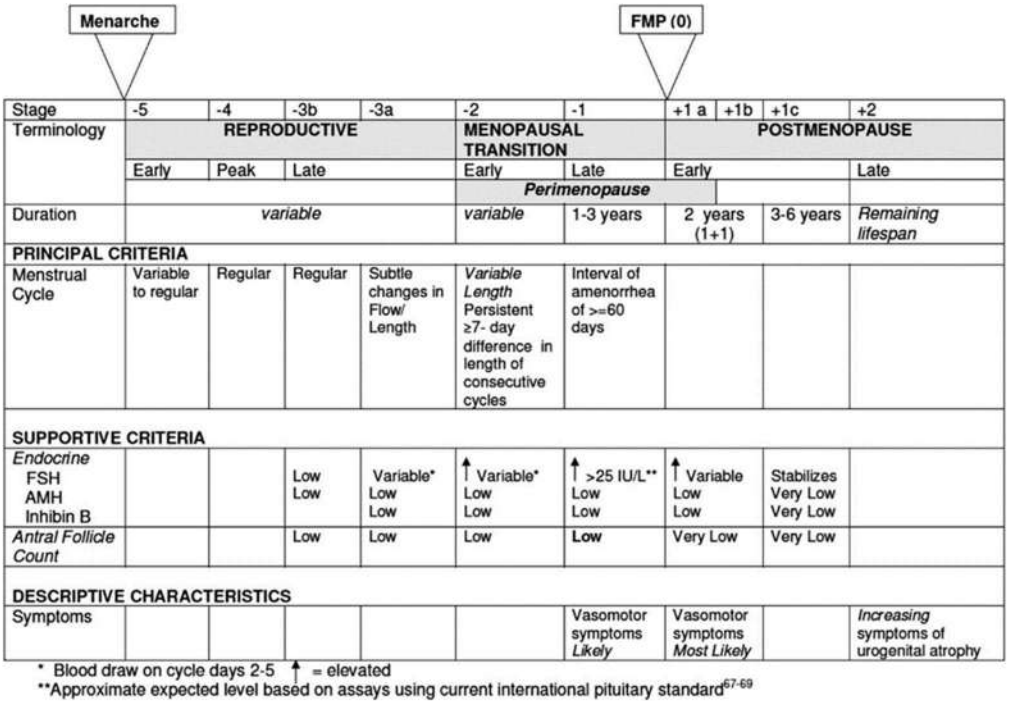
Despite the existence of an established nomenclature that facilitated a scientific consensus to describe female reproductive aging, there was still a lack of clear and objective criteria to describe the stages of female reproductive aging, which led to the realization of the Stages of Reproductive Aging Workshop (STRAW) in 2001 [16].

The STRAW, organized by the North American Menopause Society (NAMS), aims to standardize the nomenclature of the stages of female reproductive life, and proposed the division of the climacteric into stages [19, 8]. The STRAW criteria divided the phases of female reproductive aging into seven distinct stages, focusing particularly on healthy women who were in natural menopause, considering menstrual cycles, endocrine/biochemical factors, signs/symptoms in other organ systems and anatomy for this definition. uterine/ovarian [18].

The STRAW suggested that the terms perimenopause and climacteric should be synonymous, with use restricted to patients or the general public, but not in scientific articles, as recommended by the WHO. The term climacteric is not used directly in the proposed classifications, but the application of the term is maintained due to its daily use, both by health care professionals and society.

Established in 2011, the Stages of Reproductive Aging Workshop: STRAW + 10 (Figure 1), proposed a new classification, to improve the definitions applied and communication in the field. This proposal considers the female reproductive life since menarche, highlighting three main categories: Reproductive (early, peak, and late), Menopausal Transition (early and late) and Post-Menopause (early and late). Ten stages are considered (-5. -4, -3b, -3a, -2, -1, +1a, +1b, +1c, +2), six of which (-5. -4, -3b, -3a, -2, -1) before the end of the menstrual period (FMP) and four (+1a, +1b, +1c, +2) after the end of the menstrual period. Their classification is made according to a Main Criterion and Supporting Criteria. The variability of the menstrual cycle is considered the Main Criterion for the diagnosis and classification of reproductive stages. Supporting Criteria (laboratory) are anti-Mullerian hormone (AMH) levels, inhibin B and FSH, and antral follicle count. Vasomotor symptoms and urogenital atrophy are also considered. Diagnosis and classification for healthy women is done by the Main Criterion. The symptomatology and the supporting criteria are used in the diagnosis and classification of women with polycystic ovary syndrome and primary ovarian insufficiency or who have undergone specific surgical procedures (endometrial ablation, unilateral oophorectomy or hysterectomy), which alter the menstrual cycle without determining the total depletion of ovarian hormones, as the underlying disease may be the explanation for changes in the menstrual cycle [20].

The STRAW criteria were an advance in understanding women's health and are considered the current gold standard for defining terms related to female reproductive aging [18], thus representing the core of OntoWoH.



Menarche

FMP (0)

Stage	-5	-4	-3b	-3a	-2	-1	+1 a	+1b	+1c	+2
Terminology	REPRODUCTIVE				MENOPAUSAL TRANSITION		POSTMENOPAUSE			
	Early	Peak	Late		Early	Late	Early			Late
					Perimenopause					
Duration	*variable*				*variable*	1-3 years	2 years (1+1)		3-6 years	*Remaining lifespan*
PRINCIPAL CRITERIA										
Menstrual Cycle	Variable to regular	Regular	Regular	Subtle changes in Flow/ Length	*Variable Length* Persistent ≥7- day difference in length of consecutive cycles	Interval of amenorrhea of >=60 days				
SUPPORTIVE CRITERIA										
Endocrine FSH AMH Inhibin B			Low Low	Variable* Low Low	↑ Variable* Low Low	↑ >25 IU/L** Low Low	↑ Variable Low Low		Stabilizes Very Low Very Low	
Antral Follicle Count			Low	Low	Low	Low	Very Low		Very Low	
DESCRIPTIVE CHARACTERISTICS										
Symptoms						Vasomotor symptoms *Likely*	Vasomotor symptoms *Most Likely*			*Increasing symptoms of urogenital atrophy*

* Blood draw on cycle days 2-5 ↑ = elevated

**Approximate expected level based on assays using current international pituitary standard[67-69]

Figure 1: Stages of Reproductive Aging Workshop: STRAW + 10 [20]

2.1.1. Symptomatology and Treatment

Many women go through the climacteric without complaints or medication, while others have symptoms that vary in diversity and intensity. In both cases, it is essential that the woman receives follow-up, to promote her health, obtaining an early diagnosis, immediate treatment of injuries and prevention of possible damage. In 2008, data showed that the increase in women's symptoms and problems in this period are not just endocrine events of the climacteric and menopause, but are also a reflection of social and personal circumstances, such as the personal (psychological state), marital, family and professional situation, and the decrease of endogenous estrogen, are the main factors that influence the symptoms intensity and clinical signs [13].

To define the symptoms and signs resulting from the interaction between sociocultural, psychological and endocrine factors that manifest themselves in aging women, the term Climacteric Syndrome is applied [21], whose diagnosis is based on detailed anamnesis complemented with a thorough physical examination [22, 23].

Among the most common climacteric symptoms, we mention physical and psychological discomfort, such as hot flashes in the upper part of the body, insomnia, vaginal dryness, palpitations, headaches, increased irritability, difficulty concentrating, memory failures, anxiety and depression. In the long term, there may be repercussions on the bone, cardiovascular and urinary systems [9].

The first step in the Climacteric Syndrome treatment should be to encourage healthy habits, such as physical exercise, maintaining a balanced diet and avoiding smoking, to alleviate symptoms [24].

The need for treatment for Climacteric Syndrome is based on the intensity of short-term symptoms and the risk for long-term conditions such as Osteoporosis and Cardiovascular Disease. Different therapies are used to reduce, mainly, vasomotor symptoms, without using hormones. Such therapies can be divided into non-pharmacological or behavioral therapies, pharmacological therapies (for example, antidepressants) and alternative pharmacological therapies (such as herbal medicines) [25].

Hormone therapy is considered the most effective treatment for dealing with vasomotor symptoms and Genitourinary Syndrome of Menopause and it has been prevented bone loss and fractures [26]. It should be noted that there is an opportunity window for the use of hormone therapy, which covers the first ten years from the date of the woman's last menstruation, or age between 50 and 60 years. The closer to menopause the start of hormone therapy, the greater the benefits and the lower the risks. Even within the opportunity window, it is essential that a careful and judicious analysis of a patient be carried out before the prescription of hormone therapy [27].

2.2. Methodological Aspects and Development of OntoWoH

An ontology is a formal representation of relevant concepts and relationships of a system, which allows sharing and reuse of acquired and represented knowledge [28]. They can be classified into foundational ontologies (serve as basis for the construction of other ontologies, as they describe more abstract concepts), domain (represent concepts and relations associated to a specific domain, such as Medicine), task (represent concepts and relations of tasks or generic activities that contribute to the solution of problems independent of the domain) and of application (depend on a particular domain and a specific task at the same time) [29].

For the construction of OntoWoH, classified as a domain ontology, the guidelines of the Unified Foundational Ontology (UFO) [30] were adopted, which enables a consistent integration with the methodology chosen for the development of OntoWoH (SABiO).

To deal with the different aspects of reality, UFO is divided into three categories: UFO-A (ontology of *endurants or continuants*), focused on objects related to the domain to be modeled; UFO-B (ontology of *perdurants or occurents*), focused on events and processes, that is, on the actions existing between the domain objects, and UFO-C (ontology of social aspects), which aims to define concepts of social scope [30, 31]. The development of OntoWoH required constructs from UFO-A and UFO-B.

OntoWoH was developed applying the Systematic Approach for Building Ontologies (SABiO) methodology [32]. SABiO has two groups of processes: The Development Process and the Support Process. The activities of the Development Process give rise to the artifacts: Reference Ontology, a conceptual model centered on the expressiveness of the domain, and Operational Ontology, which is a machine-processable artifact mapped from the corresponding reference ontology. Initially, the Purpose Identification and Requirements Definition and Knowledge Acquisition activities were developed to elaborate OntoWoH. Knowledge Acquisition was carried out mainly through bibliographical research, supported by domain specialists during the activity.

Figure 2 presents the SABiO flow adopted in the construction of OntoWoH, highlighting the results obtained in some activities (for example, during the evaluation, quality control was obtained through verification and validation tasks).

During the Knowledge Acquisition, it was identified that, despite some differences, there is a convergence of viewpoints and terminologies in the area of Women's Health, which contributed to the consistent choice of which concepts would be used in the project. Regarding the term Climacteric, the WHO definition was chosen, followed by STRAW conceptualization, from which the terms related to the staging of female reproductive aging were adopted.

A pending activity to be performed is Reuse. A brief simplified search showed that there was no ontology like OntoWoH, but further research is needed to confirm this statement.

The Visual Paradigm editor was chosen for modeling the ontology, due to its robustness, offering free access in its community version and for enabling integration with the OntoUML Plugin for Visual Paradigm, which provides the use of UFO stereotypes and meta-properties natively in the editor [33].

Quality control tasks has taken two manner: (1) Verification, which has been the result of tasks such as carefully reviewing graphical issues of diagrams (such as visual arrangement of symbols to aid diagram legibility and cardinality definition in proper relationships), to check the nomenclature of the represented terms (according to the domain and standard nomenclature usually adopted in conceptual models), to check that all types of entities and relationships are characterized via a UFO stereotype; (2) validation, which was the result of domain understanding tasks and checking whether such understanding was reflected in the ontology, which basically occurred through periodic discussions between the team of specialists, and was also reflected in the competence questions (CQ) elaborated (during the requirements elicitation activity) and that later had a response characterized from the navigation in the diagrams.

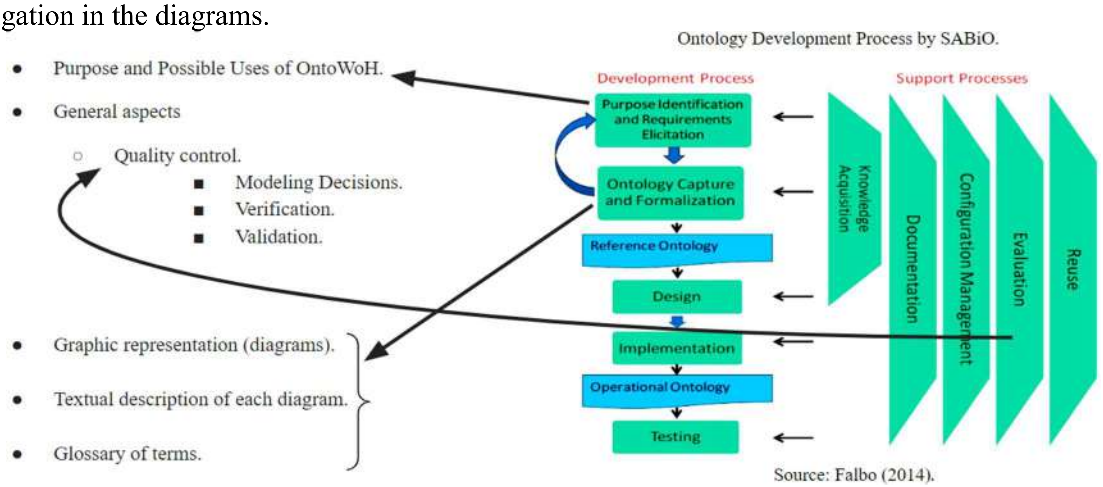


Figure 2: Application of the SABiO in the construction of OntoWoH.

Next, some CQ are presented, which helped in understanding the domain, and the traceability of these CQ in the ontology, through the reasoning adopted to answer the question from the concepts of OntoWoH.

- CQ 1: What are the particularly feminine events and at what stages of a woman's life can such events occur?
  Answer: In addition to the phases of a Person[2]'s life, during her life a Woman can go through particularly feminine events, which are the Menstrual Cycle, Pregnancy, Breastfeeding and the Climacteric phases, which are the Early and Late Menopausal Transitions and the first part of the Early Post-Menopause. The Menstrual Cycle results in Pregnancy or Menstruation, in which Menarche (first Menstruation, which usually occurs in adolescence) and Menopause (last Menstruation, which usually occurs in adulthood) stand out. The Menarche marks the beginning, and Menopause marks the end, of Menacme. During Menacme, which goes from adolescence to adulthood, Pregnancy and Breastfeeding may occur. The Climacteric phases (the Early and Late Menopausal Transition, the first part of the Early Post-Menopause) that occur before the Menopause, occur in adulthood.
- CQ 2: Which stages of women's reproductive aging are considered in STRAW + 10?
  Answer: The Female Reproductive Aging Staging System considers different phases of a woman's life, defined between the occurrence of Menarche and Menopause (central milestone). The phases are: Pre-Menopause (subdivided into Reproductive Period and Menopausal Transition), Perimenopause and Post-Menopause (subdivided into Early Post-Menopause and Late Post-Menopause). Such phases are equivalent to Stages, within which women are diagnosed and classified according to Criteria for Diagnosis and Classification (which is explained in a separate diagram) established by the Female Reproductive Aging Staging System.

[2] To facilitate the identification of the ontology concepts used in the text, the cited terms are highlighted with a different font than that used in the rest of the text.

3. Reference Ontology for Women's Health (OntoWoH)

The purpose of the Reference Ontology for Women's Health - OntoWoH - is to consistently represent the domain of Women's Health, with emphasis on the climacteric and menopause in its first version, highlighting the most susceptible symptoms in this period, possible treatments for the identified symptoms, the phases of reproductive aging, as well as essentially female characteristics and events. As for use, it is expected to be used by domain experts, technical specialists and other stakeholders as an object for communication and understanding of the domain, facilitating the dissemination of information in different segments and in addition to being used as a teaching and learning instrument for health professionals. Also, it can serve as income for Decision Support Information Systems in the area of Women's Health.

OntoWoH is a case study of the joint use of different technologies/methodologies, namely: the SABiO methodology, the UFO foundational ontology, the Visual Paradigm editor with the OntoUML plugin, the intense bibliographic research and the interaction with domain and technical experts.

Figure 3 presents an overview of the OntoWoH architecture, showing the packages and the links between them. OntoWoH has been organized into packages to facilitate its representation, understanding and use. With regard to the use of colors to highlight some property of the model, the colors defined by the VP tool for the different UFO stereotypes were preserved. To facilitate the identification of the modules and each concept of the ontology, we adopted the nomenclature in which each word (simple or compound) starts with the first letter in capital letters.

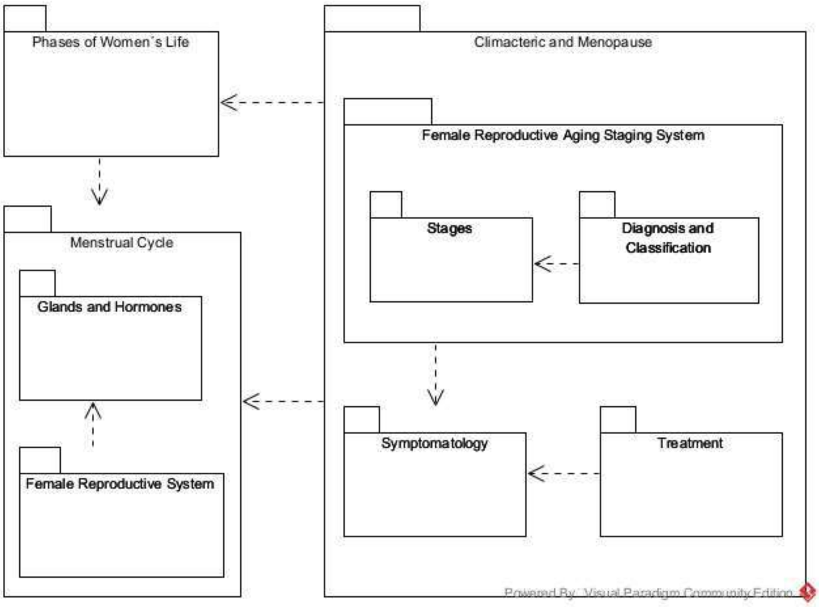

Phases of Women´s Life
Climacteric and Menopause
Female Reproductive Aging Staging System
Stages
Diagnosis and Classification
Menstrual Cycle
Glands and Hormones
Female Reproductive System
Symptomatology
Treatment
Powered By Visual Paradigm Community Edition

Figure 3: OntoWoH Architecture Overview

The Menstrual Cycle module presents particularities of the Menstrual Cycle, such as its components and symptoms of the Menstrual Syndrome and involves both the glands and hormones, as well as the female reproductive system. Its sub-modules are: (i) The Glands and Hormones, that presents types of glands, types of hormones and the main hormones relevant to the female universe; (ii) The Female Reproductive System presents the human body systems and particularities of the female reproductive system, mainly internal organs.

The Climacteric and Menopause module refers to the staging of Female Reproductive Aging, which involves the diagnosis and classification of women's situation, as well as symptomatology and treatment approaches for the climacteric phase and menopause. Its sub-modules are: (i) The Symptomatology sub-module presents early and late manifestations, transient and non-transient, resulting from female reproductive aging, associated with menopause; (ii) The Treatment sub-module presents the types of approaches for treating women who suffer from the Climacteric Syndrome; (iii) The Female Reproductive Aging Staging System sub-module presents the female reproductive aging staging system according to the STRAW + 10 criteria, which is in turn divided into: (iii.1) The Diagnosis and

Classification sub-module presents the criteria diagnosis and classification of women according to the STRAW + 10; (iii.2) The Stages sub-module presents the stages that make up the STRAW + 10.

Finally, the Phases of a Woman's Life module presents the phases of a woman's life and the main events in the female universe, under the aspects of her reproductive aging, focusing on menopause.

The OntoWoH 1.0 modules have nine diagrams (its graphical representation), a complementary descriptions and a glossary of terms for the applied concepts. In this paper some of these diagrams are presented, offering an overview of the main concepts adopted. An example of how the glossary of terms is organized[3]: Concept: Menstruation. Definition: Menstruation is the shedding of the lining of the uterus (endometrium) accompanied by bleeding. It occurs in approximately monthly cycles throughout a woman's reproductive life, except during pregnancy. Menstruation starts during puberty (at menarche) and stops permanently at menopause. Source: https://www.msdmanuals.com/pt-br/casa/problemas-de-sa%C3%BAde-feminina/biologia-do-sistema-reprodutor-feminino/ciclo-menstrual (original definition in Portuguese). Diagram(s): Phases of a Woman's Life Module, Menstrual Cycle Diagram.

We can observe in the presented diagrams the use of several UFO stereotypes, which served as guidelines for some modeling decisions that occurred throughout the construction of OntoWoH. Among the UFO stereotypes applied, we mention kind, which encompasses individuals who have the principle of identity and are existentially independent - as is the example of `Person` (each person is unique and does not depend on the existence of another concept to exist) and `phase` (as in woman´s life phases), whose instantiation is determined by an inherent property of the individual, and the change between phases occurs due to changes in values in the intrinsic properties of the instances (for example, John instantiates the `Child` phase if his age is less than 12 years old). A property is called quality when it has a measurable value in its quality dimensions (for example, a person's weight or age) and is called a mode if it cannot be represented by a measurement system (for example, a person's headache) [30][4].

In the phases of a Woman's Life diagram (Figure 4) we visualize female events according to the phases in which they occur, culminating in `Menopause`, which occurs due to `Ovarian Failure` and can happen naturally or unnaturally.

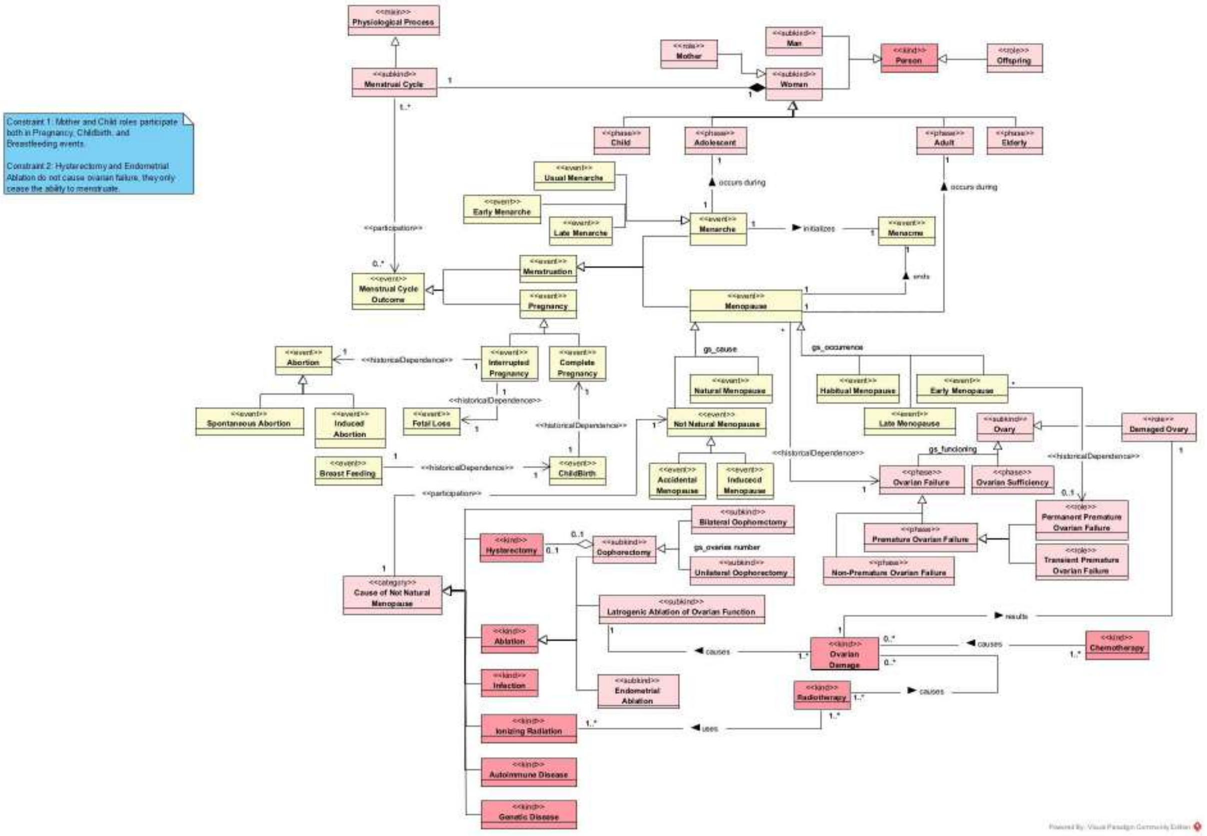
During her life, a `Woman` usually goes through a reproductive period, a non-reproductive period and a transition period from the reproductive to the non-reproductive period, which is represented in the diagram of the `Female Reproductive Aging Staging System` (Figure 5) - the core of OntoWoH, which considers different phases of a woman's life, defined from the occurrence of `Menarche` and `Menopause` (central milestone).

The usual phases of woman´s life are `Pre-Menopause` (divided into the `Reproductive Period` and `Menopausal Transition`), `Perimenopause` and `Post-Menopause` (specialized in `Early Post-Menopause` and `Late Post-Menopause`). Such phases are equivalent to `Stages`, within which women are diagnosed and classified according to `Criteria for Diagnosis and Classification` established by the `Female Reproductive Aging Staging System`.

`Menarche` and `Menopause` are `Milestones` in a `Woman`'s life. `Menarche` marks the beginning and `Menopause` (which occurs in `Perimenopause`) marks the end of the `Reproductive Period` (classified into `Early Reproductive Period, Early Reproductive Period` and `Late Reproductive Period`, which marks the moment when fertility begins to decline and during which the woman begins to notice changes in her menstrual cycles). The `Menopausal Transition`, which corresponds to the period of time characterized by the onset of endocrine, biological and clinical changes that indicate the approach of `Menopause`, is subdivided into `Early Menopausal Transition` (marked by increased variability in the duration of the menstrual cycle) and `Late Menopausal Transition`, marked by the occurrence of amenorrhea for 60 days or more.

[3] For access to the complete glossary of terms, as well as the complete ontology specification, please contact the authors of the research.

[4] For more details on the UFO foundation, please consult publications that discuss such artifact specifically, as in [30] and [31].

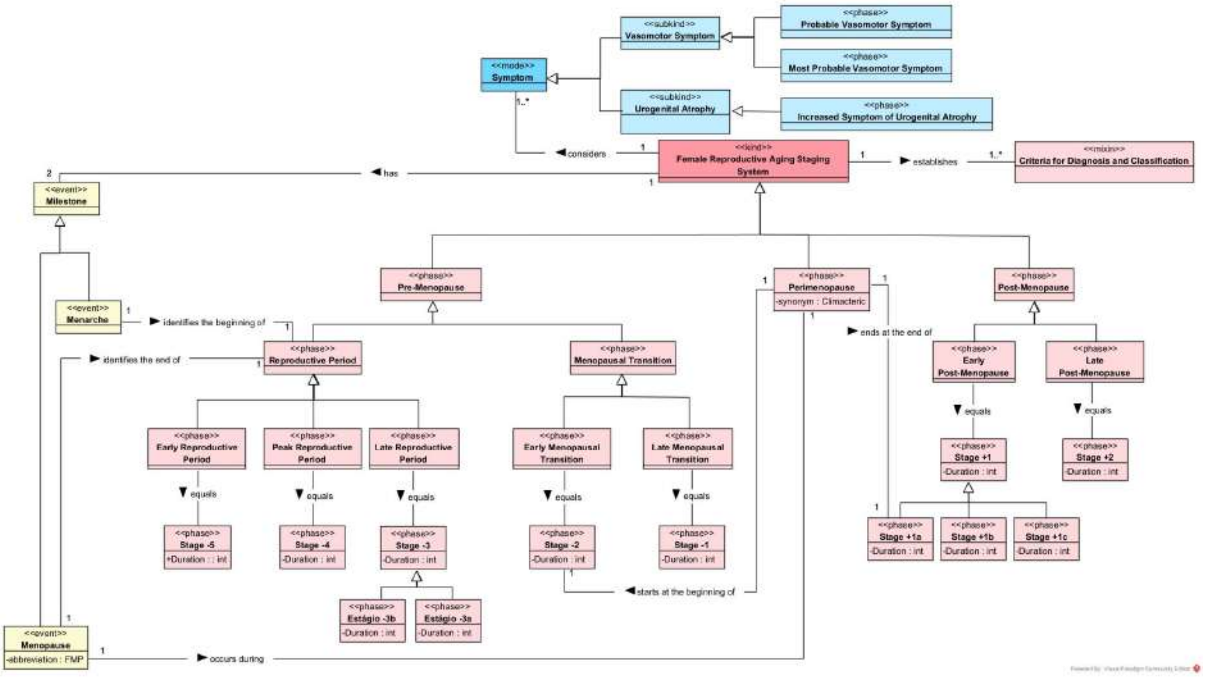


<<mixin>> Physiological Process
<<subkind>> Menstrual Cycle
<<role>> Mother
<<subkind>> Man
<<subkind>> Woman
<<kind>> Person
<<role>> Offspring
<<phase>> Child
<<phase>> Adolescent
<<phase>> Adult
<<phase>> Elderly
<<event>> Usual Menarche
<<event>> Early Menarche
<<event>> Late Menarche
<<event>> Menarche
occurs during
initializes
<<event>> Menacme
ends
<<participation>>
<<event>> Menstrual Cycle Outcome
<<event>> Menstruation
<<event>> Pregnancy
<<event>> Menopause
gs_cause
gs_occurrence
<<event>> Abortion
<<historicalDependence>>
<<event>> Interrupted Pregnancy
<<event>> Complete Pregnancy
<<event>> Spontaneous Abortion
<<event>> Induced Abortion
<<event>> Fetal Loss
<<event>> Breast Feeding
<<event>> ChildBirth
<<event>> Natural Menopause
<<event>> Not Natural Menopause
<<event>> Accidental Menopause
<<event>> Induced Menopause
<<event>> Habitual Menopause
<<event>> Early Menopause
<<event>> Late Menopause
<<subkind>> Ovary
<<role>> Damaged Ovary
gs_functioning
<<phase>> Ovarian Failure
<<phase>> Ovarian Sufficiency
<<role>> Permanent Premature Ovarian Failure
<<phase>> Premature Ovarian Failure
<<role>> Transient Premature Ovarian Failure
<<phase>> Non-Premature Ovarian Failure
<<subkind>> Bilateral Oophorectomy
<<kind>> Hysterectomy
<<subkind>> Oophorectomy
gs_ovaries number
<<subkind>> Unilateral Oophorectomy
<<category>> Cause of Not Natural Menopause
<<subkind>> Latrogenic Ablation of Ovarian Function
<<kind>> Ablation
causes
<<kind>> Ovarian Damage
results
<<kind>> Chemotherapy
<<kind>> Infection
<<subkind>> Endometrial Ablation
<<kind>> Radiotherapy
<<kind>> Ionizing Radiation
uses
<<kind>> Autoimmune Disease
<<kind>> Genetic Disease
1
1..*
0..*
0..1
*
Powered By: Visual Paradigm Community Edition

Constraint 1: Mother and Child roles participate both in Pregnancy, Childbirth, and Breastfeeding events.

Constraint 2: Hysterectomy and Endometrial Ablation do not cause ovarian failure, they only cease the ability to menstruate.

Figure 4: Woman's Life Phases diagram



<<mode>> Symptom
<<subkind>> Vasomotor Symptom
<<phase>> Probable Vasomotor Symptom
<<phase>> Most Probable Vasomotor Symptom
<<subkind>> Urogenital Atrophy
<<phase>> Increased Symptom of Urogenital Atrophy
considers
<<kind>> Female Reproductive Aging Staging System
establishes
<<mixin>> Criteria for Diagnosis and Classification
has
<<event>> Milestone
<<event>> Menarche
identifies the beginning of
identifies the end of
<<phase>> Pre-Menopause
<<phase>> Reproductive Period
<<phase>> Menopausal Transition
<<phase>> Perimenopause
-synonym : Climacteric
<<phase>> Post-Menopause
ends at the end of
<<phase>> Early Post-Menopause
<<phase>> Late Post-Menopause
<<phase>> Early Reproductive Period
<<phase>> Peak Reproductive Period
<<phase>> Late Reproductive Period
<<phase>> Early Menopausal Transition
<<phase>> Late Menopausal Transition
equals
<<phase>> Stage -5
+Duration : : int
<<phase>> Stage -4
-Duration : int
<<phase>> Stage -3
-Duration : int
<<phase>> Stage -2
-Duration : int
<<phase>> Stage -1
-Duration : int
<<phase>> Stage +1
-Duration : int
<<phase>> Stage +2
-Duration : int
<<phase>> Stage +1a
-Duration : int
<<phase>> Stage +1b
-Duration : int
<<phase>> Stage +1c
-Duration : int
<<phase>> Estágio -3b
-Duration : int
<<phase>> Estágio -3a
-Duration : int
starts at the beginning of
<<event>> Menopause
-abbreviation : FMP
occurs during
1
2
1..*
Powered By: Visual Paradigm Community Edition

Figure 5: Female Reproductive Aging Staging System Diagram

`Perimenopause` begins in the first phase of the `Menopausal  Transition` and ends in `Early Post-Menopause`, more specifically in `Stage +1a`, which lasts one year, and ends when FSH and estradiol levels stabilize and marks the end of twelve months period of amenorrhea necessary to define the occurrence of PMF. Symptoms, particularly vasomotor ones, are more likely to occur during this stage.

4. Final Considerations

The benefits of integrating the areas of Computer Science and Health Care are varied, which justifies the growing number of HCIS's that aim to improve the quality of Health Care. Ontologies, such as OntoWoH, are tools used to overcome some HCIS challenges (such as the integration of large volumes of data and knowledge sharing), which allows Health Care professionals to perform more accurate analyzes, to have greater access to knowledge and means to offer better service to the population, whose portion directly benefited with OntoWoH are women. In these cases, the ontology is an initial artifact for the development of a technological framework. On the other hand, an ontology is also a final artifact, when applied as a tool for teaching, learning and communication between stakeholders in the domain.

The research carried out showed the complexity and richness of information in the specific domain of Climacteric and Menopause. Using STRAW and domain specialists as main source of information, OntoWoH can be considered as an artifact to clarify and establish a terminology and to help in the semantic interoperability of the domain. We tried to represent it in a clear and reliable manner, but there are many possibilities for future work concerning the domain. Among these possibilities, we cite specific causes of Amenorrhea, adverse effects of therapeutic approaches, and complementary exams indicated for climacteric women.

Also, regarding STRAW + 10, there are some issues that are not included in version 1.0 of OntoWoH, but will be considered in the future: the need for operationalization and transparent reporting of pre-menopause, which is currently not explicitly stated; to establish, as minimum requirement, that regular menstruation be defined as the number of menstrual cycles in a time interval of at least three months and the need to consider the usefulness of introducing normative age ranges as a supplementary criterion for defining the stages of female reproductive aging [18].

OntoWoH was conceived with the aim of arousing greater interest and knowledge for the stage of a woman's life that is menopause, in different environments and groups, as well as facilitating the dissemination of information. We highlight among the purposes and possible uses of OntoWoH, the need to develop a HCIS with its database based on OntoWoH, whose users would be Health Care professionals, mainly.

To prepare OntoWoH for the next step, according to SABiO, one (or more than one) operational ontology must be developed. It is in our plans to continue the process, mapping the OntoWoH reference ontology (based on UFO) to, for example, an operational ontology developed in the OWL language. This mapping will open up a series of possibilities, mainly to explore reuse and integration with existing ontologies, such as those available in OBO Foundry. It is expected that a step like this will cause the revision of the reference ontology itself, to better explore reuse situations.

As lessons learned from the project, the importance of involving domain experts in the ontology development activities is highlighted, in order to obtain results that better reflect the reality and, at the same time, make them defenders of the use of ontologies [34].

OntoWoH is open to improvement and expansion, as well as reuse, including an integration to OntoSaúde [35] – what has already been started, and application as basis to a technological tool. An initial quality control was carried out by both technical experts (through technical reviews) and its likely future users (through domain experts with whom the model was discussed), and so its first version was considered ready to be released[5]. New activities, such as empirical studies, are desired to ensure the quality of OntoWoH. In summary, the soap opera OntoWoH has begun, but there are still many chapters to go, maybe even some spin-off.

5. References

[1] J. Mylopoulos, “Conceptual Modelling and Telos”, In P. Loucopoulos, & R. Zicari (Eds.), Conceptual Modelling, Databases and CASE: An Integrated View of Information Systems Development, New York: Wiley, 1992.

[5] At this moment, the ontology and associated documentation are not available for download on the Internet. However, for possible interested parties, the authors are open to contact and to send the material developed.

[2] W. M. C. Medeiros, SISOnt: istema de informação em saúde baseado em ontologias, Dissertação de Mestrado, Universidade Federal do Rio Grande do Norte, 2009.

[3] F. B. Nardon, Compartilhamento de Conhecimento em Saúde Utilizando Ontologias e Bancos de Dados Dedutivos, 2003, Tese de Doutorado, Universidade de São Paulo.

[4] S. Isotani, I. I. Bittencourt, Dados abertos conectados: em busca da web do conhecimento, Novatec Editora, 2015.

[5] S. Coelho; Y. F. Porto, Saúde da Mulher, 2 ed. Belo Horizonte, Nescon, UFMG, 2013.

[6] F. Chaimowicz et al, Saúde do idoso, Belo Horizonte: Nescon/Coopmed, 2009.

[7] F. Angum et al, The Prevalence of Autoimmune Disorders in Women: A Narrative Review, Cureus v. 12,5 e8094, 13 May 2020. doi:10.7759/cureus.8094

[8] Federação Brasileira das Associações de Ginecologia e Obstetrícia (FEBRASGO), Climatério Manual de Orientação, São Paulo, 2010.

[9] M. D. H. A. Da Rocha, P. A. Da Rocha, Do climatério à menopausa. Revista científica do ITPAC, v. 3, n. 1, 2010.

[10] BRASIL, Ministério da Saúde, Política Nacional de Atenção Integral à Saúde da Mulher: princípios e diretrizes, Brasília, 2011.

[11] L. D. C. Silva, M. V. Mamede, Prevalence and severity of menopausal symptoms in women with coronary artery disease/Prevalência e intensidade de sintomas climatéricos em mulheres com doença arterial coronariana, Revista de Pesquisa Cuidado é Fundamental Online, v. 12, 2020, p. 305-312, doi: 10.9789/2175-5361.rpcfo.v12.6755

[12] Nations, United et al, World population ageing 2019 highlights, United Nations, 2019.

[13] BRASIL, Ministério da Saúde. Secretaria de Atenção à Saúde, Departamento de Ações Programáticas Estratégicas, Manual de Atenção à Mulher no Climatério / Menopausa, Série A, Normas e Manuais Técnicos Série Direitos Sexuais e Direitos Reprodutivos – Caderno, n.9, Brasília, 2008.

[14] Word Health Organization. Menopause. 2022, Oct, https://www.who.int/news-room/fact-sheets/detail/menopause

[15] R. E. Jones, K. H. Lopez, Human reproductive biology, Academic Press, 2013.

[16] Janet E. Hall, Endocrinology of the Menopause, Endocrinology and Metabolism Clinics of North Americ, 2015 Sep;44(3):485-96m doi: 10.1016/j.ecl.2015.05.010. PMID: 26316238; PMCID: PMC6983294.

[17] W. H. Utian, The International Menopause menopause-related terminology definitions, Climacteric, v. 2, n. 4, 1999. p. 284-286.

[18] A. Ambikairajah, E. Walsh., N. Cherbuin, A review of menopause nomenclature, Reproductive Health, v. 19, n. 1, p. 1-15, 2022. doi:10.1186/s12978-022-01336-7

[19] M. R. Soules et al, Executive summary: stages of reproductive aging workshop (STRAW), Fertility and Sterility, v. 76, n. 5, november 2001. doi.org/10.1016/S0015-0282(01)02909-0

[20] S. D. Harlow et al., Executive summary of the Stages of Reproductive Aging Workshop + 10: addressing the unfinished agenda of staging reproductive aging, The Journal of Clinical Endocrinology & Metabolism, Volume 97, Issue 4, 1 april 2012, p. 1159–1168. doi:10.1210/jc.2011-3362.

[21] W. H. Utian, Ovarian function, therapy-oriented definition of menopause and climacteric, Maturitas, v. 1, n. 22, p. 65, 1995.

[22] BRASIL, Ministério da Saúde, Instituto Sírio-Libanês de Ensino e Pesquisa, Protocolos da atenção básica: saúde das mulheres, Brasília, 2016.

[23] L. F. C. Baccaro et al., Initial evaluation in the climacteric. Revista Brasileira de Ginecologia e Obstetrícia, v. 44, p. 548-556, 2022.

[24] L. F. C. Baccaro, Conversando sobre o climatério. Entrevista concedida à Letícia Martins, Femina, Federação Brasileira das Associações de Ginecologia e Obstetrícia (FEBRASGO), São Paulo, v. 50, n. 5, 2022. p. 272-74.

[25] L. H. S. C. Paiva, A. L. R. Valadares, L. F. C. Baccaro, Como tratar os sintomas vasomotores sem o emprego da terapêutica hormonal?, In: L. M. Pompei et al., Consenso Brasileiro de Terapêutica Hormonal da Menopausa, Associação Brasileira de Climatério (SOBRAC), São Paulo, Leitura Médica, 2018, p. 147-153.

[26] North American Menopause Society et al., The 2022 hormone therapy position statement of The North American Menopause Society. Menopause: The Journal of The North American Menopause Society, Vol. 29, No. 7, may 2, 2022. p. 767-794, DOI: 10.1097/GME.0000000000002028

[27] F. Guidozzi et al., South African Menopause Society revised consensus position statement on menopausal hormone therapy, 2014. SAMJ: South African Medical Journal, v. 104, n. 8, p. 537-543, 2014.

[28] N. Guarino, D, Oberle, S. Staab, What is an Ontology? In: S. Staab, R. Studer (Eds.), Handbook on Ontologies, Second ed., p.1–20, 2009.

[29] G. Guizzardi, Ontological Foundations for Structural Conceptual Models, Tese de doutorado, Universidade de Twente, 2005.

[30] R. A. Falbo, SABiO: Systematic approach for building ontologies, CEUR Workshop Proceedings, 1301, 2014.

[31] OntoUML VP Plugin, github.com/OntoUML/ontouml-vp-plugin.

[32] N. Guarino, Formal Ontology in Information Systems, Itália: IOS Press, 1998.

[33] G. Guizzardi, R. A. Falbo and R. S. S. Guizzardi, "The role of Foundational Ontologies for Domain Ontology Engineering: a case study in the Software Process Domain," in IEEE Latin America Transactions, vol. 6, no. 3, pp. 244-251, July 2008, doi: 10.1109/TLA.2008.4653854.

[34] E. Norris et al., (2021), Why and How to Engage Expert Stakeholders in Ontology Development: Insights From Social and Behavioural Sciences. Journal of Biomedical Semantics, 12(4). doi:10.1186/s13326-021-00240-6

[35] D. R. Costa, M. G. S. Teixeira, S. D. Rissino, T. S. Guarnier, O Uso da Abordagem SABiO na Construção do Overview de OntoSaúde. In J. P. A. Almeida, & G. Guizzardi (Eds.), Engineering Ontologies and Ontologies for Engineering: Celebrating Ricardo Falbos's Career, Vitória, ES, Brasil, 2020, pp. 82 – 98.

Using ontology embeddings with deep learning architectures to improve prediction of ontology concepts from literature

Pratik Devkota[1], Somya D. Mohanty[2] and Prashanti Manda[1]

[1]*Informatics and Analytics, University of North Carolina at Greensboro*
[2]*United Healthcare*

Abstract

Natural language processing methods powered by deep learning have been well-studied over the past years for the task of automated ontology-based annotation of scientific literature. Many of these approaches focus solely on learning associations between text and ontology concepts and use that to annotate new text. However, a great deal of information is embedded in the ontology structure and semantics. Here, we present deep learning architectures that learn not only associations between text and ontology concepts but also the structure of the ontology. Our experiments show that creating architectures that are capable of learning the structure of the ontology result in enhanced annotation performance.

Keywords

natural language processing, gene ontology, deep learning, ontology annotation, ontology embeddings

1. Introduction

Biological ontologies are widely used for representing biological knowledge across a wide range of sub-domains ranging from gene function to clinical diagnoses to evolutionary phenotypes [1, 2, 3]. While the ontologies provide the necessary structure and concepts, the real benefits of the ontologies can be reaped only when knowledge in scientific literature is represented using these ontologies through annotation. The scale and pace of scientific publishing demands sophisticated, fast, and most importantly, automated ways of processing scientific literature to annotate relevant pieces of text with ontology concepts [4].

Natural Language Processing (NLP) techniques beginning with lexical analysis, standard machine learning approaches, and of late, powered by deep learning models have made big strides in this area [5, 6, 7, 8, 9, 10]. Most NLP approaches for automated ontology annotation treat the task as that of named entity recognition where relevant entities are identified and associated with snippets of text. However, ontology based annotation is different from named entity recognition in that there is a great amount of information embedded in the structure and

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

p_devkota@uncg.edu (P. Devkota); mohanty.somya@gmail.com (S. D. Mohanty); p_manda@uncg.edu (P. Manda)
0000-0001-5161-0798 (P. Devkota); 0000-0002-4253-5201 (S. D. Mohanty); 0000-0002-7162-7770 (P. Manda)

semantics of an ontology whereas generic entities can be independent objects. Knowledge of the ontological structure and relationships is a crucial part of biological annotation when performed by a human curator. It is therefore imperative to develop NLP models that are cognizant of the ontological hierarchy and can effectively incorporate it into the prediction mechanism for improved ontology concept recognition.

The automated annotation models previously developed by this team [11, 8, 12, 10, 9] have shown good accuracy in recognizing ontology concepts from text. In these studies our focus was to teach the models to learn associations between text and ontology concepts found in the gold standard corpus and use that knowledge to create new annotations. In a few studies, we experimented with different techniques of using the ontology structure as one of the inputs in a bid to improve annotation performance [10, 8, 9]. In some cases, these systems are able to predict the same ontology concept as the ground truth in the gold standard data achieving perfect accuracy. Incorporating ontology structure was a bid to improving partial accuracy in cases where the model does not achieve a perfect match to the actual annotation. Our hypothesis was that having knowledge of the ontology structure would enable the model to choose a closely related/semantically similar concept to the actual annotation thereby improving overall annotation performance as evaluated by semantic similarity.

Our goal in this study is to develop deep learning architectures that learn not only patterns in text but also the ontology structure. Our hypothesis is that the process of learning the ontology structure would in turn improve prediction of annotations. Deep learning models learn patterns in text and annotations from a gold standard corpus and similarly, we need to provide a gold standard representation of the ontology structure so the models can learn to predict the ontology structure.

In this study, we use graph embeddings for representing the ontology structure. These graph embeddings are used as a reference and reinforcement tool for the model as it learns to predicts the ontology structure. Semantic embedding of large knowledge graphs has been long used successfully for predictive tasks including natural language processing [13]. In recent years, these semantic embeddings have been extended to OWL ontologies resulting in approaches that can create embeddings for ontology concepts that effectively represent the structure and semantics of the ontology [13, 14]. These embedding algorithms translate ontologies represented as directed acyclic graphs into a vector space where the structure and the inherent semantics of the graph are preserved [15].

There are several approaches for learning ontology embeddings [16, 17] each with different strengths. The approaches differ based on whether the ontology is directed, weighted, if it dynamically changes over time, and the approach for learning the network [17]. In this study, we selected Node2Vec [14] for learning ontology embeddings from the Gene Ontology since it is widely used in literature for this task [17].

We use the Colorado Richly Annotated Full Text Corpus (CRAFT) as a gold standard for training and testing the performance of our architectures [18]. CRAFT is a widely used training resource for automated annotation approaches. The current version of the CRAFT corpus (v4.0.1) provides annotations for 97 biological/biomedical articles with concepts from 7 ontologies including the GO.

We hypothesize that the added information gained from ontology embeddings can improve model performance in recognizing ontology concepts from scientific literature. We persent two deep learning architectures and explore how the different architectures combined with the inclusion of ontology embeddings impacts annotation performance.

2. Related Work

The rise of deep learning in the areas of image and speech recognition has translated into text-based problems as well. Preliminary research has shown that deep learning methods result in greater accuracy for text-based tasks including identifying ontology concepts in text [5, 19, 20, 21, 8]. These methods use vector representations that enable them to capture dependencies and relationships between words using enriched representations of character and word embeddings from training data [7].

Our initial foray into this area involved a feasibility study of using deep learning for the task of recognizing ontology concepts [11]. In a comparison of Gated Recurrent Units (GRUs), Long Short Term Memory (LSTM), Recurrent Neural Networks (RNNs), and Multi Layer Perceptrons (MLPs) along with a new deep learning model/architecture based on combining multiple GRUs, we found GRUs to outperform the rest. These findings indicated that deep learning algorithms are a promising avenue to be explored for automated ontology-based curation of data.

In 2020, we presented new architectures based on GRUs and LSTMs combined with different input encoding formats for automated annotation of ontology concepts [8]. We also created multi level deep learning models designed to incorporate ontology hierarchy into the prediction. Surprisingly, inclusion of ontology semantics via subsumption reasoning yielded modest performance improvement [8]. This result indicated that more sophisticated approaches to take advantage of the ontology hierarchy are needed.

Continuing this work, a 2022 study [12] presented state of the art deep learning architectures based on GRUs for annotating text with ontology concepts. We augmented the models with additional information sources including NCBI's BioThesauraus and Unified Medical Language System (UMLS) to augment information from CRAFT for increasing prediction accuracy. We demonstrated that augmenting the model with additional input pipelines can substantially enhance prediction performance.

Our next work explored a different approach to providing the ontology as input to the deep learning model [8]. Subsequently, we presented an intelligent annotation system [10] that uses the ontology hierarchy for training and predicting ontology concepts for pieces of text. Here, we used a vector of semantic similarity scores to the ground truth and all ancestors in the ontology to train the model. This representation allowed the model to identify the target GO term followed by "similar" GO terms that are partially accurate predictions. We showed that our ontology aware models can result in a 2% - 10% improvement over a baseline model that doesn't use ontology hierarchies.

Our most recent contribution presented a method called Ontology Boosting [9]. A key component of this approach is to combine the prediction of the deep learning architectures with the graph structure of ontological concepts. Boosting amplifies the predicted probabilities of certain concept predictions by combining them with the model predictions of the candidate's ancestors/subsumers. Results showed that the boosting step can result in a substantial bump in prediction accuracy.

3. Methods

3.1. Generating ontology embeddings

We used the Node2Vec approach for generating ontology embeddings from the Gene Ontology.

The Node2Vec algorithm consists of two steps:

1. Conduct random walks from a graph or ontology to generate sentences which are a list of ontology concepts. Once all random walks are conducted, the set of all sentences makes the corpus which is a representation of the ontology.
2. The Word2Vec [22] algorithm is applied on the corpus to learn and generate embeddings for each concept in the ontology. These embeddings are low dimensional vector representations of ontology concepts.

These embeddings or feature vectors can be used in downstream tasks such as classification or natural language processing. in a downstream task such as node classification.

We set the weight of all edges to 1 for weighted random walks indicating that all edges are weighted equally. The length of random walk was set to 5 and the walk number set to 100. Dimensionality of embeddings was set to 128 and batch size is set to 50 and the model was trained for 2 epochs to learn the embeddings.

3.2. Deep Learning Architectures

Here, we present and test three sets of architectures:

1. Baseline
   - Tag only (*TO*)
   - Ontology Embedding only (*OEO*)
2. Cross-connected:
   - Tag to Ontology Embedding ($T- > OE$)
   - Ontology Embedding to Tag ($OE- > T$)
3. Multi-connected:
   - Embedding to Tag to Embedding

3.2.1. Baseline Architectures

We created two baseline architectures (Figure 1): Tag only (*TO*) and Ontology Embedding only (*OEO*). The *TO* architecture predicts tags/ontology IDs while the *OEO* architecture predicts ontology embeddings. The *TO* architecture has previously been presented in our prior work [10]. This architecture has been adjusted to create the *OEO* architecture that predicts ontology embeddings. Both baseline architectures consist of input pipelines, embedding/latent representations, and a deep learning model and produce either a probability vector of ontology IDs (*TO*) or an ontology embedding (*OEO*) as the output.

The baseline architectures use three inputs. Each word in a sentence from the CRAFT corpus is represented by three inputs - 1) token (X_{train}^{token}), 2) character sequence (X_{train}^{char}), 3) Parts-Of-Speech (POS) (X_{train}^{POS}).The token (X_{train}^{token}) input, is a sequential tensor consisting tokens, each represented with a high dimensional one hot encoded vector. The character sequence (X^{char}_{train}) is also a sequential tensor consisting of character sequences present in a word/token. POS tags (X^{POS}_{train}) indicate the type of words in a sentence.

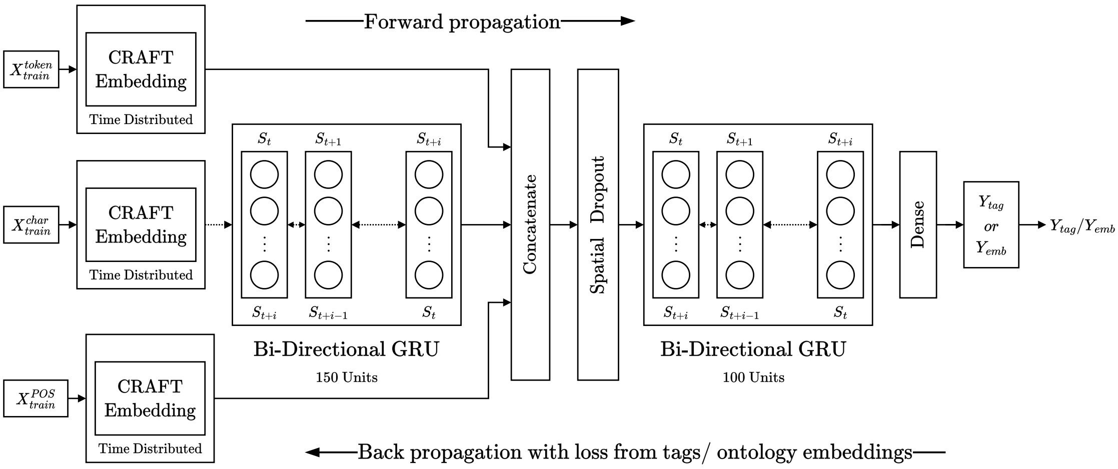


Figure 1: Tag-only (*TO*) and Ontology Embeddings only (*OEO*) baseline architectures. *TO* produces a tag/ontology ID (Y_{tag}) as output in the final block whereas *OEO* produces an ontology embedding (Y_{emb}). The architectures also differ in that is used during back propagation to compute gradient loss. *TO* uses tags while *OEO* uses Node2Vec ontology embeddings.

Embeddings are used to provide a compressed latent space representation for very high dimensional input components. For example, the one hot vectorization of an individual word has a dimensionality of 34,166 (vocabulary size). In order to represent them succinctly and with contextual representation, we use supervised embeddings created from the CRAFT corpus. Note that these embeddings are different from the ontology embeddings discussed above. These embeddings provide low dimensional representations of words in the training corpus and do not use the ontology in any way.

Both baseline architectures use a bi-directional gated recurrent model (Bi-GRU). The choice of Bi-GRU for the architectures was informed by several of our prior works where this model has consistently outperformed other models such as CNNs, RNNS, and LSTMs [8, 11]. Architecture hyper-parameters were evaluated using a grid search approach. We used Adam [23] as our optimiser for all of the experiments with a default learning rate of 0.001.

The two baseline architectures differ primarily because of what they produce as output and what they use during the propagation stages. In *TO*, the output is a tag/ontology ID where each word in the input data is mapped to either a GO annotation or a non-annotation. *TO* takes the hidden/learned representations of the input from the preceding layers of the network and applies softmax activation to produce a probability distribution over all possible ontology ids. The predicted vector output values and ground truth values are compared to compute sparse categorical cross entropy as loss, followed by backpropagation which involves computing the gradients of the loss with respect to the model's weights. The ontology ID with the highest probability is regarded as the prediction.

In contrast, *OEO* uses ground truth ontology embeddings generated using Node2Vec during back propagation and for computing the loss functions. The intuition is that providing ontology embeddings to the architecture during the propagation stages will enable it to get an understanding of the ontology structure and eventually enable it to make more accurate and intelligent predictions. The output of *OEO* is an ontology embedding. The predicted ontology embedding is compared to all ground truth ontology embeddings using cosine similarity calculation. The ground truth ontology embedding that is most similar to the predicted embedding is identified and the ontology ID associated with it is treated as the architecture's prediction. Accuracy metrics are then computed by comparing the predicted ontology ID to that in the CRAFT corpus.

3.3. Cross connected architectures

We developed two cross connected architectures: 1) Tag to Ontology Embedding ($T- > OE$) and 2) Ontology Embedding to Tag ($OE- > T$). Here we test if connecting the tag and ontology embedding architectures causing one to inform the prediction of the other would result in improved accuracy and if the direction of the connection matters. The $T- > OE$ architecture (Figure 2) has two different outputs, tags/ontology ids and ontology embeddings. The tag output (Y_{tag} in Figure 2) is concatenated with the output of the main Bi-GRU layer to give a higher dimensional vector output. The concatenation is then passed through dense layers to further learn the hierarchical representations of the ontology before generating an ontology embedding for each input token. This predicted ontology embedding is compared with the ground truth ontology embeddings learned using Node2Vec. Using cosine similarity as the loss function, loss is calculated and the gradients are backpropagated to adjust the model's weight for convergence.

In $OE- > T$, the ontology embedding output (Y_{emb} is concatenated with the output of the main Bi-GRU layer to give a higher dimensional vector output. The concatenation is then passed through dense layers before generating a tag for each input token. This predicted tag is compared with the ground truth tag in CRAFT. The loss is calculated and the gradients are backpropagated to adjust the model's weight for convergence. The $OE- > T$ architecture can be depicted by switching the Y_tag and Y_emb blocks as well as the two outputs in Figure 2.

Figure 3 presents an explanation of the $T- > OE$ cross-connected architecture on three example tokens. Cross connected architectures differ from the baseline architectures by producing both tags and ontology embeddings instead of one or the other. Here, we show that the training/ inference is done on a sequence of tokens "vesicle", "formation", and "in" (which are parts of a sentence in the CRAFT corpus) as it is evaluated by the network. Each token is preprocessed to obtain the representative tensors – X^{token}_{train}, X^{char}_{train}, X^{POS}_{train} which are passed through embedding layers learned from CRAFT. The embedding of X^{char}_{train} is also passed via a Bi-GRU layer. All of the resulting values are concatenated to be processed via the main Bi-GRU layer. The output from 'Tag Dense Layer' is concatenated with the output of main 'Bi-GRU layer' and passed as input to the 'Ontology Embedding Dense Layer' where the model generates ontology embeddings for each of the input tokens.

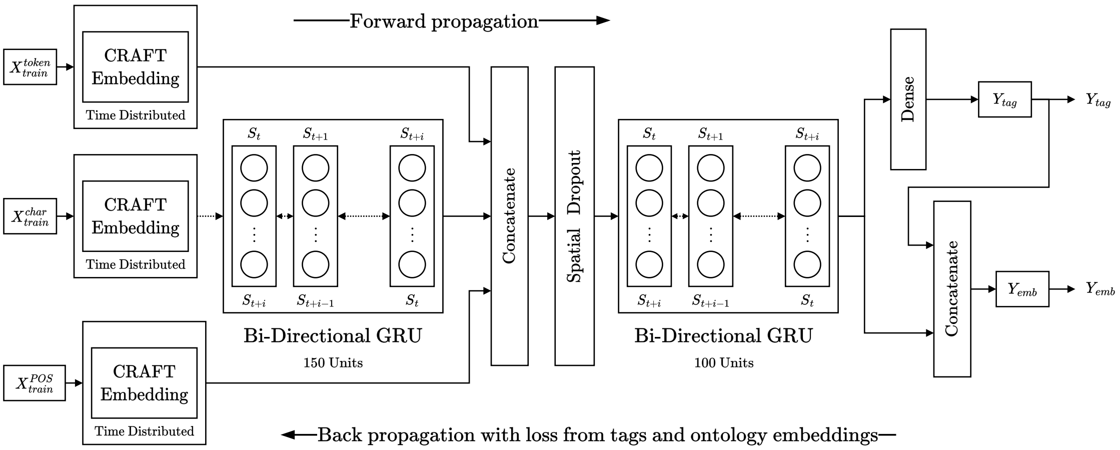


Figure 2: Tag to Embedding architecture ($T- > OE$). The tag output is further fed to the ontology embedding prediction block resulting in a better embedding prediction.

3.4. Multi-connected Architecture

The final architecture ($OE- > T- > OE$) explores if ontology embeddings can be improved iteratively by connecting a preliminary ontology embedding output to the tag output enabling improvements to the tag prediction. This predicted tag block is connected back to the ontology embedding block to urge further learning.

3.5. Performance Evaluation Metrics

We evaluate our architectures using a modified F1 score and semantic similarity [24]. Metrics such as F1 are designed for traditional information retrieval systems that either retrieve a piece of information or fail to do so (a binary evaluation). However, this is not a true indication of the performance of ontology-based retrieval or prediction systems where the notion of partial accuracy applies. A model might not predict the exact concept as a gold standard but might predict the parent or an ancestor of the ground truth as indicated by the ontology. Semantic similarity metrics [24] designed to measure different degrees of similarity between ontology concepts can be leveraged to measure the similarity between the predicted concept and the actual annotation to quantify the partial prediction accuracy. Here, we use Jaccard similarity [24] that measures the ontological distance between two concepts to assess partial similarity.

Since the majority of tags in the training corpus are non-annotations, the model predicts them with great accuracy. In order to avoid biasing the F1 score, we omit accurate predictions of non-annotations and focus instead on annotations only report a relatively conservative modified F1 score.

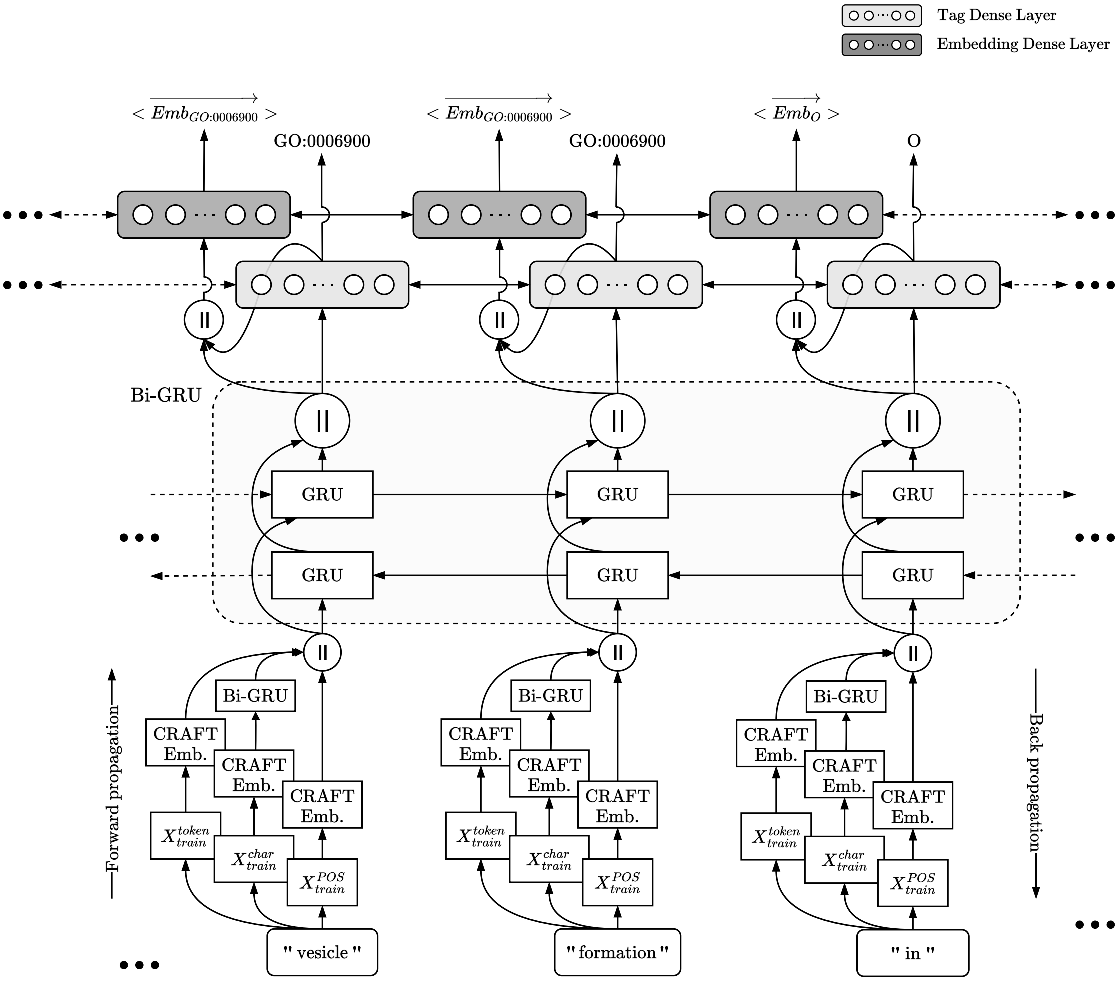


Figure 3: Illustration of the working of the $T->OE$ architecture with an example sequence. The architecture produces two outputs - 1) a tag and 2) an ontology embedding.

4. Results and Discussion

The CRAFT v4.0.1 dataset contains 18689 annotations pertaining to 974 concepts from the three GO sub-ontologies across 97 articles. Table 1 provides further information of the coverage of GO terms in CRAFT.

The baseline tag-only architecture (*TO*) resulted in a 0.80 F1 and a 0.83 semantic similarity score. The baseline ontology embeddings only architecture (*OEO*) resulted in a 0.65 F1 and a 0.74 semantic similarity.

Among the two cross-connected architectures, we found that the Tag to Ontology Embedding architecture ($T->OE$) substantially outperformed the $OE->T$ architecture according to F1 and was able to achieve similar performance as measured by semantic similarity. This indicates that $T->OE$ is better at generating exactly matching predictions resulting in high F1 and semantic similarity. In contrast, $OE- > T$ performs better are generating semantically similar matches rather than exact matches leading to lower F1 than semantic similarity scores.

The $T- > OE$ architecture was able to improve upon OEO's prediction of ontology embeddings by 23% (F1) and 9.4% (semantic similarity). We observed relatively modest improvements to TO's tag prediction with 3.8% (F1) and 1.2% (semantic similarity).

Connecting ontology embedding output to the tag output ($OE- > T$) either did not improve on the embedding prediction (F1) or resulted in a slight improvement (semantic similarity). $OE- > T$ did produce improvements for tag prediction over the TO model by 3.7% (F1) and 1.2% (semantic similarity). The multi-connected architecture did poorly in comparison to the cross-connected architectures.

Overall, the results suggest that architectures that use ontology embeddings only without learning associations between text and annotations perform poorly. The other takeaway is that connecting tag predictions to the ontology embedding block ($T- > OE$) and letting embedding prediction learn from the predicted tag iteratively results in more robust architectures. The $T- > OE$ cross-connected architecture results in improved performance in predicting both tags and ontology embeddings across both metrics.

Table 1
Coverage of GO ontology concepts and annotations in the CRAFT corpus

GO sub-ontology	Concepts in ontology	Total annotations in CRAFT	Unique occurences in CRAFT
Biological Process (BP)	30490	18392	710
Cellular Component (CC)	4463	6976	241
Molecular Function (MF)	12257	464	5

Table 2
Performance metrics of the three sets of architectures measured by F1 and Jaccard semantic similarity

Architecture	**Ontology Embedding F1 Score**	**Ontology Embedding Similarity Score**	**Tag F1 Score**	**Tag Similarity Score**
Baseline Architectures				
Tag-only (TO)	-	-	0.80	0.83
Ontology Embedding Only (OEO)	0.65	0.74	-	-
Cross-connected Architectures				
Tag to Ontology Embedding ($T- > OE$)	**0.80**	**0.81**	**0.83**	0.84
Ontology Embedding to Tag ($OE- > T$)	0.64	0.75	0.83	**0.84**
Multi-connected Architecture				
$OE- > T- > OE$	0.78	0.80	0.82	0.83

Acknowledgments

This work is funded by a CAREER grant to Manda from the Division of Biological Infrastructure at the National Science Foundation of United States of America (#1942727).

References

[1] T. R. Dalmer, R. D. Clugston, Gene ontology enrichment analysis of congenital diaphragmatic hernia-associated genes, Pediatric research 85 (2019) 13–19.

[2] D. Lee, N. de Keizer, F. Lau, R. Cornet, Literature review of snomed ct use, Journal of the American Medical Informatics Association 21 (2014) e11–e19.

[3] R. C. Edmunds, B. Su, J. P. Balhoff, B. F. Eames, W. M. Dahdul, H. Lapp, J. G. Lundberg, T. J. Vision, R. A. Dunham, P. M. Mabee, et al., Phenoscape: identifying candidate genes for evolutionary phenotypes, Molecular biology and evolution 33 (2015) 13–24.

[4] W. Dahdul, T. A. Dececchi, N. Ibrahim, H. Lapp, P. Mabee, Moving the mountain: analysis of the effort required to transform comparative anatomy into computable anatomy, Database 2015 (2015).

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).

[6] M. R. Boguslav, N. D. Hailu, M. Bada, W. A. Baumgartner, L. E. Hunter, Concept recognition as a machine translation problem, BMC bioinformatics 22 (2021) 1–39.

[7] M. A. Casteleiro, G. Demetriou, W. Read, M. J. F. Prieto, N. Maroto, D. M. Fernandez, G. Nenadic, J. Klein, J. Keane, R. Stevens, Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature, Journal of biomedical semantics 9 (2018) 13.

[8] P. Manda, S. SayedAhmed, S. D. Mohanty, Automated ontology-based annotation of scientific literature using deep learning, in: Proceedings of The International Workshop on Semantic Big Data, SBD '20, Association for Computing Machinery, New York, NY, USA, 2020. URL: https://doi.org/10.1145/3391274.3393636. doi:10.1145/3391274.3393636.

[9] P. Devkota, S. D. Mohanty, P. Manda, Ontology-powered boosting for improved recognition of ontology concepts from biological literature (2023).

[10] P. Devkota, S. Mohanty, P. Manda, Knowledge of the ancestors: Intelligent ontology-aware annotation of biological literature using semantic similarity, Proceedings of the International Conference on Biomedical Ontology (2022).

[11] P. Manda, L. Beasley, S. Mohanty, Taking a dive: Experiments in deep learning for automatic ontology-based annotation of scientific literature, Proceedings of the International Conference on Biomedical Ontology (2018).

[12] P. Devkota, S. D. Mohanty, P. Manda, A gated recurrent unit based architecture for recognizing ontology concepts from biological literature, BioData Mining 15 (2022) 1–23.

[13] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, I. Horrocks, Owl2vec*: Embedding of owl ontologies, Machine Learning 110 (2021) 1813–1845.

[14] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: Proceedings

of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 855–864.

[15] M. Ou, P. Cui, J. Pei, Z. Zhang, W. Zhu, Asymmetric transitivity preserving graph embedding, in: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016, pp. 1105–1114.

[16] H. Cai, V. W. Zheng, K. C.-C. Chang, A comprehensive survey of graph embedding: Problems, techniques, and applications, IEEE transactions on knowledge and data engineering 30 (2018) 1616–1637.

[17] I. Makarov, D. Kiselev, N. Nikitinsky, L. Subelj, Survey on graph embeddings and their applications to machine learning problems on graphs, PeerJ Computer Science 7 (2021) e357.

[18] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, L. E. Hunter, Concept annotation in the craft corpus, BMC Bioinformatics 13 (2012) 161. URL: https://doi.org/10.1186/1471-2105-13-161. doi:10.1186/1471-2105-13-161.

[19] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (2017) i37–i48.

[20] C. Lyu, B. Chen, Y. Ren, D. Ji, Long short-term memory rnn for biomedical named entity recognition, BMC bioinformatics 18 (2017) 462.

[21] X. Wang, Y. Zhang, X. Ren, Y. Zhang, M. Zitnik, J. Shang, C. Langlotz, J. Han, Cross-type biomedical named entity recognition with deep multi-task learning, arXiv preprint arXiv:1801.09851 (2018).

[22] K. W. Church, Word2vec, Natural Language Engineering 23 (2017) 155–162.

[23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

[24] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, F. M. Couto, Semantic similarity in biomedical ontologies, PLoS computational biology 5 (2009).

KOSonto: An ontology for knowledge organization systems, their constituents, and their referents

Jean Noel Nikiema[1,2,*], Fleur Mougin[3], Vianney Jouhet[4,3] and Stefan Schulz[5,6]

[1] *Department of Management, Evaluation and Health Policy, School of Public Health, Université de Montréal, Canada*
[2] *Centre de recherche en santé publique, Université de Montréal et CIUSSS du Centre-Sud-de-l'Île-de-Montréal, Canada*
[3] *Univ. Bordeaux, Inserm UMR 1219, Bordeaux Population Health Research Center, team AHeaD, Bordeaux, France*
[4] *CHU de Bordeaux, Pôle de santé publique, Service d'information médicale, Bordeaux, France*
[5] *Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, Graz, Austria*
[6] *Averbis GmbH, Freiburg, Germany*

Abstract

The structure of knowledge organization systems (KOSs) – domain vocabularies, thesauri, terminologies, classification systems, and ontologies – follows different architectural principles and semantic theories. However, many use cases require their integrated use in a given domain. Building a common framework for KOSs is then a prerequisite for any principled account of their use when data annotated by different KOSs should be integrated. We propose an approach rooted in formal ontology, the aim of which is to harmonize the description of the domain itself with the description of the representational artifacts that claim to organize and represent knowledge of this domain. We propose a transparent framework for describing KOSs with a focus on the biomedical domain. Using comprehensive and consistent terminology, we formalize what KOSs represent by introducing KOSonto, an ontology that characterizes representational artifacts on the one hand and describes the relationships to their referents in the domain of application on the other hand. KOSonto uses OWL-DL axioms and is built under BFO and IAO. It accounts for a range of elements that are characteristic of different kinds of KOSs. We illustrate how KOSonto can be used to describe typical biomedical KOSs such as ICD-10, SNOMED CT, and MeSH. Further work will improve the alignment of KOSonto to foundational ontologies and apply this framework to optimize the creation, use, and reuse of mappings between heterogeneous KOSs.

Keywords

Knowledge organization systems, formal ontologies, KOSonto

1. Introduction

For decades, biomedical informatics has focused on artifacts for organizing domain knowledge [1]. Terminologies, controlled vocabularies, dictionaries, thesauri, classifications, nomenclatures, and ontologies – for which we will use the overarching term "**knowledge organization systems**" (KOSs) [2] – vary in scope, granularity, and design principles. Most of them

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*] Corresponding author.

jean.nikiema@umontreal.ca (J. N. Nikiema); fleur.mougin@u-bordeaux.fr (F. Mougin); vianney.jouhet@chu-bordeaux.fr (V. Jouhet); stefan.schulz@medunigraz.at (S. Schulz)
0000-0002-9396-6423 (J. N. Nikiema); 00000-0002-7436-3010 (F. Mougin); 0000-0001-5272-2265 (V. Jouhet); 0000-0001-7222-3287 (S. Schulz)

were created to address specific use cases, often without making their model of meaning explicit. Efforts invested in KOSs have contributed to interoperability and to a better understanding of (classes of) domain entities and the terminological units that refer to them. However, many of the current denominations for KOS are imprecise and misleading. Particularly, the terms "Controlled vocabulary" and "Terminology systems" are misleading because they suggest that these systems mostly describe human language, although they are often applied to KOSs that have a clear focus on the description of domain entities. For example, systems like ICD-10, ChEBI, the Gene Ontology, and the NCBI taxonomy are often referred to as controlled vocabularies although they do not convey any lexical information. Words such as "term", "concept", "ontology", "entity", "descriptor", "class", "property", and "relation" are used inconsistently, which leads to misunderstandings, particularly in cross-disciplinary cooperations. It also raises the question of why attempts to standardize the entities of a given domain, *viz.* biomedicine and health, are not underpinned by a meta-level standardization of the description of the realm of representation itself, *viz.* its symbols and its relation to the language and the reality of the domain itself.

Only a few *principled analyses* of the domain of representation itself have been carried out. In the 1990s, MIVoc set out to standardize basic semantic notions in medical informatics [3]. Since then, library science, linguistics, philosophy, and the semantic web have fueled knowledge organization activities, without much uptake of MIVoc, which was withdrawn in 2006. Addressing the need for international cooperation, a multilingual scale standardization initiative has been proposed [4], again with only minor generalizations of the usage of this initiative. The result is a confusing mix of approaches, theories, technical terms, and conceptualizations that have been waiting for a thorough cleanup.

It is imperative to carry out this long-awaited cleaning-up process due to the growing importance of diverse KOSs, such as ICD-10, SNOMED CT, MeSH, and MedDRA for biomedical knowledge management. Each of these systems possesses a particular structure, user community, scenarios of use, and representational philosophy. This is particularly pressing where the integration and alignment of KOSs require semantic harmonization between KOSs (and data annotated with them). Indeed, KOSs have always been used in combination. Therefore, despite their structural and semantic heterogeneity, semantic links and correspondences between their elements are sought. There is a long tradition of building bridges between biomedical KOSs. KOS alignment has been an important topic in the semantic web and knowledge graph communities. Numerous heuristics for KOS alignment/mapping/harmonization have been described [5]. In the biomedical domain, the UMLS Metathesaurus [6] has been of great benefit as the longest-running and most enduring effort of KOSs alignment. Additional efforts such as HeTOP [7] and BioPortal [8] are also noteworthy. These resources, driven by practical needs, support some integration of KOSs, without emphasizing their semantic particularities [9]. Only the UMLS Metathesaurus mappings are continually revised manually by domain experts, although there are some automation initiatives [10].

This paper proposes a ontology-based framework for describing KOSs themselves, together with *what they denote.* We propose "KOSonto", an OWL model under Basic Formal Ontology

(BFO) 2020[1] [11] and using elements of the Information Artifact Ontology (IAO)[2] [12], which is available at https://github.com/JeanNikiema/kosonto. We believe that only after a principled ontological analysis of the constituents of typical KOSs and their representational commitments the value and restrictions of different kinds of KOSs will be sufficiently determined, and the consequences of an alignment between KOSs of different kinds be predicted. A clear and consistent terminology for KOSs themselves should avoid the pitfalls of divergent interpretations of ill-defined words like "ontology", "concept", "entity", "property", or "knowledge". We deliberately avoid most of these words (or only use them in a clear context, such as "SNOMED CT concept").

The paper is structured as follows: Section 2 presents KOSonto based on a KOS content framework; in Section 3, we describe some biomedical KOSs (ICD-10, MeSH, SNOMED CT, and a small HL7 value set) according to KOSonto; and we discuss our main findings in Section 4.

2. KOSonto – The ontology of knowledge organization systems

2.1. Kinds and content of knowledge organization systems

Different elements support the characterization of KOSs as a meaningful ontological category. First, we consider all KOSs **information content entities** (ICEs) according to IAO. ICEs are immaterial but inherent in one or more material bearers [13]. For instance, the content of the KOS (e.g., SNOMED CT) just as of a work of fiction (e.g., Victor Hugo's "Les Contemplations"), both ICEs, can be stored in many electronic storage systems at the same time. In addition, the constituents of KOSs are also ICEs, e.g., the concept 195967001 – "Asthma (disorder)" or the poem "Vere novo", respectively. Secondly, KOSs and their constituents are linked to referents (detailed in the following subsection) in a real or fictional world they intend to represent. Finally, they are artifacts and as such created by humans. With these three characteristics, KOSs are **representational artifacts** [14] constituted by a number of **representational units** (**RUs**), which denote particular **referents**. Accordingly, a backbone of a common KOS framework requires a clear characterization of both referents and RUs. KOSonto uses OWL-DL axioms and is classifiable using the HermiT reasoner [15]. The reason for the choice of OWL-DL is popularity, tool support, and human understandability, but also the appropriateness of this language for representing a mostly static domain. KOSonto includes a typology of possible referents of RUs and their ontological foundation, the ontological nature of the RUs themselves, the symbols used in KOSs, and a typology of KOSs.

2.2. The referent as the kind of entity of what is represented in KOSs

A referent is **what** is represented in a KOS, more precisely the thing in the world that is denoted by a RU of a KOS. Everything can be a referent, as the only requirement for being a referent is **to be denoted**. "Referent" is therefore not a meaningful ontological category and not represented in KOSonto. KOSOnto is based on three fundamental categories: particular entities, type entities, and class entities.

[1] https://basic-formal-ontology.org/BFO-2020/

[2] https://github.com/information-artifact-ontology/IAO/

Particular entities[3] are concrete in space and time, have objective existence and ontological significance, and exist independently of human perception or language [16]. We introduce the class *Particular* as the disjunction of *bfo:Continuant* and *bfo:Occurrent.*

Type entities (or **types**) correspond to repeatable, or instantiable (often qualified as abstract) entities. When a type is instantiated by a specific particular, this particular can be referred to as an instance of this type. We divide them further into:

- *Universals as defined by Aristotle*: encompass anything that can be instantiated by particulars. Aristotelian universals are *immanent*, i.e. they exist in their instances, which precludes universals without instances such as unicorns or intergalactic travels.
- *Types by intension*: represent entities of meaning given by means of a formal definition, comparable to the characteristic function in set theory. They do not necessarily extend to things in reality. They allow for defining, e.g., a unicorn as being a pink horse with a single horn, without however claiming its existence. Intensional meanings have classes of particulars as their extensions, including empty classes.
- *Types by extension*: depend on their particular members, without further descriptions. For example, the set {America, Europe, Africa, Asia, Antarctica, Oceania} necessarily and sufficiently corresponds to what is understood by "Continent". However, such representations are not very common in the biomedical domain.
- *Cognitive types*: correspond to mental representations rooted in language and sensory perceptions, regardless of any concrete correlation. A cognitive representation of a particular or conceptual unicorn or the use of the word "centaur" does not mean that unicorns or centaurs exist. We introduce *FictionalType* as a subclass of *CognitiveType* for those things that exist in a fictional world only.

Class entities (or **classes**) are central elements of description logics and are sometimes considered equivalent to types [17]. Their set-theoretic semantics emphasizes the importance of classes of particulars. This is why we grant them a prominent status as siblings of *Particular entities* and *Type entities* and fully define them as the extension of a type that can have only particulars as members:

$$\textit{Class}\ \textbf{equivalentTo}\ (\textbf{extends}\ \textit{some type}\ )\ \text{and}\ (\textbf{has\_member}\ \textit{only Particular}). \tag{1}$$

In KOSonto, classes are always implemented as classes of OWL particulars, ensuring a coherent and well-defined framework. It manages to sidestep the logical conundrum presented by Russell's paradox [18] and maintains its operational efficacy. Thus, the definition of classes is provided by the set of characteristics of their actual or potential members, the undefined cognitive meaning, or the universal properties inherent in all their members. Whether a class is *currently* or *supposed to be* empty is not a unifying criterion. For example, the classes that extend the types *Unicorn*, *Centaur*, and *Elf* are not identical. Indeed, classes may exist without any particular member having all the characteristics identified to be a member. Classes are *defined* if their necessary and sufficient characteristics allow a particular entity to be recognized as a member of this class; otherwise, the classes are considered *primitive*.

[3] To avoid confusion we highlight that particular entities are modelled as OWL individuals (A-Box elements), but types are also modelled as A-Box entities.

Summing up, types have actual or at least hypothetical instances; the instances of a type are the members of the class it extends to. In OWL, the operator to express class membership is, rather confusingly, **rdf:type**[4]. OWL requires a bi-partition between classes (T-box entities) and individuals (A-box entities). For understanding our model, it is therefore important to be aware of the ontological notion of particular as introduced above and the technical notion of "OWL individuals". Note that all *types* in the ontological sense are therefore modelled in KOSOnto as *OWL individuals*, along with particulars proper. KOSonto introduces the object property **instance_of** as the relation between a particular and a type in the above sense, and **is_a** as the transitive relation between two types (modelled as A-box entities, i.e. OWL individuals). Our example ontology illustrates the parallelism between types and classes as follows. The axiom asserting that a particular that instantiates a type T_1 also instantiates T_2 if T_1 **is_a** T_2 is expressed by the property inclusion "**instance_of** ∘ **is_a** subPropertyOf **instance_of**".
We have the A-Box entities (OWL individuals) **Horse**$_{type}$, **Vertebrate**$_{type}$ and **Animal**$_{type}$ as members of *AristotelianUniversal*, on which an OWL reasoner (with A-box reasoning) computes the following expected inferences[5]:

Statements on OWL individuals		**Reasoner inference**	
Bucephalus; Facts: **instance_of Horse**$_{type}$	(2)	**Bucephalus**; Facts: **instance_of Vertebrate**$_{type}$	(5)
Horse$_{type}$; Facts: **is_a Vertebrate**$_{type}$	(3)	**Bucephalus**; Facts: **instance_of Animal**$_{type}$	(6)
Vertebrate$_{type}$; Facts: **is_a Animal**$_{type}$	(4)	**Horse**$_{type}$; Facts: **is_a Animal**$_{type}$	(7)

We then define OWL classes (T-box entities) based on the hierarchy modeled in the A-box:

Statement		**Reasoner inference**	
Horse equivalentTo **instance_of** value **Horse**$_{type}$	(8)	*Horse* subClassOf *Animal*	(11)
Vertebrate equivalentTo **instance_of** value **Vertebrate**$_{type}$	(9)	*Vertebrate* subClassOf *Animal*	(12)
Animal equivalentTo **instance_of** value **Animal**$_{type}$	(10)		

Thus, we represent the hierarchical structure of the ontology at the A-box level underneath the *type* hierarchy. In KOSonto, we introduce *FictionalType* subClassOf *CognitiveType*. While *FictionalType* does not require specifying the kind of instances of fictional types, the *FictionalType* class only allows instances of the *InformationContentEntity* (ICE) type. We can therefore distinguish between types of particulars: (i) those that are not ICEs, (ii) those that are ICEs, and (iii) those that are uncommitted: e.g., horses, centaurs (mythical human-horse hybrids), and sumxus (animals of which science disagrees whether only mythical or really existing), or green horses (potential future breeding result). At the A-box level, these entities can nevertheless be linked together, using formal relations such as **is_a** or **instance_of** but also by informal relations such as **is_narrower_than** in addition to the aforementioned relations:

Individual: **Sumxus**$_{type}$; Facts: **is_a Vertebrate**$_{type}$ (13)

Individual: **GreenHorse**$_{type}$; Facts: **is_a Horse**$_{type}$ (14)

[4] http://www.w3.org/1999/02/22-rdf-syntax-ns#type

[5] In the equations, classes are depicted in italics, KOSonto relations and A-Box entities are represented in bold, and other parts of the OWL syntax are shown in normal font. Names of OWL individuals that symbolize types have "type" as subscript.

$$\text{Individual: }\mathbf{Centaur}_{type}\text{;} \quad \text{Facts: }\mathbf{is\_narrower\_than\ Vertebrate}_{type} \tag{15}$$

$$\text{Individual: }\mathbf{Chiron}\text{;} \quad \text{Facts: }\mathbf{instance\_of\ Centaur}_{type} \tag{16}$$

with **Centaur**$_{type}$ being a member of *FictionalType* while **Sumxus**$_{type}$ or **GreenHorse**$_{type}$ could just be members of *CognitiveType* or even *TypeByIntension*. The latter case applies to the scenario where sufficient defining criteria exist, as in the case of the green horse:

$$\mathbf{GreenHorse}\text{ equivalentTo }\mathbf{instance\_of\ GreenHorse}_{type} \tag{17}$$

$$\mathbf{GreenHorse}\text{ equivalentTo }\mathbf{Horse}\text{ and }\mathbf{has\_proper\_part}\text{ some (}\mathbf{Hair}\text{ and }\mathbf{bearer\_of\ GreenColor}\text{)} \tag{18}$$

GreenHorse$_{type}$ is a member of *type*, without commitment to the existence of non-informational instances in reality, other than if it were introduced as a member of *AristotelianUniversal*. It is therefore left open whether the defined class (18) has members. This degressive description is crucial, considering that in certain cases, fictional entities, metaphors, or analogies may be introduced in KOSs to model specific conditions or processes and facilitate the understanding of intricate phenomena. The important concept “Chi” in traditional medicine systems illustrates the need for such a specification. Other examples are mental disorders whose understanding evolves over time in line with new research, clinical insights, and social acceptance (e.g., malleus maleficarum or neurasthenia) or whose existence are denied.

One might argue that types, in the broadest sense, also include relational objects such as predicates or relations, just like mathematical objects in general. A discussion of this is beyond the scope of this paper, in which we refer only to ICEs (cf. subsection 2.3).

2.3. The representational unit as an atomic representation in KOSs

Having presented the range of possible referents for KOS elements – with a digression beyond realism in order to demonstrate the model’s flexibility – we now turn to the RUs themselves and their grounding in KOSonto. All RUs are ICEs in the sense of IAO, i.e. generically dependent continuants. What can be considered an atomic form of representation in KOSs varies. We propose a different kind of RUs by using the referent as support of categorization. By answering the question from the perspective of a KOS builder (*Once we have identified the referent, how can we represent it in a KOS?*) and from a KOS user (*For a specific referent, what is its atomic form of representation in a KOS?*), we can identify *three overlapping and inclusive levels* of RUs. Each of these atomic representations can point to a referent. The minimal requirements for RUs are expressed by (19). This corresponds to **level 1**, with a term being a human-readable word or phrase, belonging to a domain-specific vocabulary. If it acts as a preferred label, it must be unique in the KOS. Labels are often artificially constructed (e.g., “Biopsy of head and neck structure”) but self-explanatory and unambiguous, regardless of their use in human communication. Although a label can act as a unique identifier, alphanumeric identifiers are more common, apart from the preferred label.

$$\begin{gathered}\textit{RepresentationalUnit}\ \text{equivalentTo}\ \textit{InformationContentEntity}\ \text{and} \\ \textbf{proper\_part\_of}\ \text{some}\ \textit{KnowledgeOrganizationSystem}\ \text{and} \\ \text{not}\ (\textbf{has\_proper\_part}\ \text{some}\ \textit{RepresentationalUnit})\ \text{and} \\ ((\textbf{has\_proper\_part}\ \text{some}\ (\textit{Literal}\ \text{and}\ \textbf{bearer\_of}\ \text{some}\ \textit{IdentifierRole})\ \text{and} \\ \textbf{has\_proper\_part}\ \text{some}\ (\textit{NaturalLanguageTerm}\ \text{and}\ \textbf{bearer\_of}\ \text{some}\ \textit{PreferredLabelRole}))\ \text{or} \\ (\textbf{has\_proper\_part}\ \text{some}\ \textit{OWL\_ClassExpression}))\end{gathered}\tag{19}$$

In most cases, KOSs offer more than one term per RU (**level 2**), and they play different roles. KOSonto distinguishes *PreferredLabelRole* from *EntryTermRole* with the subclasses *ExactSynonymRole*, *CloseSynonymRole*, *AmbiguousSynonymRole*, *HyponymRole*, *EllipticSynonymRole*. Exact synonyms have the same meaning as the preferred label, and close synonyms have a very similar meaning in the context of the use of this RU. Ambiguous synonyms belong to more than one RU in a KOS, e.g., "lead" for an electric contact or for the chemical element "Pb".

According to the definition provided in [19] for **composite representations**, KOSonto introduces three composite representations: definitions, descriptions, and exemplifications. Definitions provide sufficient and necessary criteria whereas descriptions (also known as elucidations, e.g., in BFO 2020 [11]) provide only necessary criteria. Exemplifications are descriptions by means of concrete examples. RUs with composite representations are introduced in KOSonto as *ExplainedRepresentationalUnit*:

$$\begin{gathered}\textit{ExplainedRepresentationalUnit}\ \ \text{equivalentTo}\ \ \textit{RepresentationalUnit}\ \text{and} \\ \textbf{bearer\_of}\ \text{some}\ \textit{CompositeTextualRepresentationRole}\ \text{and} \\ \textbf{has\_proper\_part}\ \text{some}\ \textit{Literal}\end{gathered}\tag{20}$$

Another form of representation is an axiomatic representation. RUs may have composite representations as axioms described in a formal language:

$$\begin{gathered}\textit{FormalRepresentationalUnit}\ \ \text{equivalentTo}\ \ \textit{RepresentationalUnit}\ \text{and} \\ \textbf{bearer\_of}\ \text{some}\ \textit{DefiniendumRole}\ \text{and} \\ \textbf{proper\_part\_of}\ \text{some}\ \textit{LogicalAxiom}\end{gathered}\tag{21}$$

Logical axioms are constituted by logical constructors (symbols and literals) respecting a specific syntax and grammar, e.g., OWL syntax. Logical axioms can be the only atomic representation available for a referent (**level 1**), or provide, with or without textual composite representations, additional information regarding an RU's referent(s) (**level 3**).

We have excluded from our analysis those KOS components that denote relational entities, i.e. in their broadest sense n-ary predicates. Whether they are "first-class" RUs or mere connectors between RUs is controversial. KOSonto includes them as subclasses of *BinaryPredicate* and *TernaryPredicate* and further elaborates on subclasses of these property classes in terms of hierarchy-building predicates, ontological predicates [20], and predicates according to domain / range restrictions in terms of types, particulars, or literals. For example, OWL object properties (*OWL_ObjectProperty*) are ontological relations that hold between particulars, and OWL datatype properties (*OWL_DataTypeProperty*) between particulars and literals. Ternary relations are not supported by OWL [21] and rarely occur in ontologies, with BFO 2020 being a notable exception [11].

3. Application to known biomedical KOSs

In this section, we briefly apply our framework to some reference biomedical KOSs.

ICD-10 is based on a strictly tree-shaped **is_narrower_than** hierarchy[6]. The disjointness of sibling RUs is a fundamental paradigm, expressed by **is_disjoint_with**. Exceptions are RUs named "others", which can be logically described as the complement of the union of their siblings. With existing clinical conditions as their referents, ICD-10 RUs can be described as denoting Aristotelian universals, apart from some examples of epistemic intrusions, such as H40.0 – "Glaucoma suspect". Such ICD-10 RUs could be seen as instantiating *CognitiveType* or alternatively *ICE* (denoting practitioner's knowledge about a patient). Finally, ICD-10 exhibits its own type-to-type relation named "exclusion", which restricts the meaning of a given RU (e.g., *Diabetes mellitus* excludes known cases of *Diabetes in pregnancy*). The knowledge about a patient's condition, as a relevant coding criterion, sheds light on the unclear nature of the referents (diseases, signs, symptoms, or diagnoses) [22]. Finally, ICD-10 exhibits terms described by *EllipticSynonymRole*, i.e. terms that implicitly require their hypernym to be human-understandable. An example is "Lip" as a label for D10.0, which is contextualized by the parent RU D10 – "Benign neoplasm of mouth and pharynx" so that human users intuitively interpret D10.0 as "Benign neoplasm of the lip".

MeSH has an informal tree structure, which can best be represented by the **is_narrower_than** relation – narrower in meaning as defined by SKOS – because the hierarchy encompasses both taxonomic and mereological aspects. Its RUs represent topics in biomedical publications, which are best interpreted as instances of *CognitiveType*. In the context of MeSH trees, RUs have a tree number as an identifier (ID). However, the same term may belong to several trees with different identifiers. Tree-related RUs and tree-independent RUs have to be distinguished. The latter ones (characterized by a separate unique identifier, or UID) can be interpreted as the hypernyms of the former ones. The hierarchical structure of MeSH is therefore more than just the overlay of trees because there is no transitivity between branches of superposed trees via a shared descriptor. As descriptors are ambiguous, there are no unique labels. For example, the descriptor "Nose" has the tree IDs A01.456.505.733, A04.531, and A09.531, as well as the UID: D009666. The coverage of entry terms and free-text definitions is large. Due to the limited granularity of MeSH, many entry terms are not synonyms but hyponyms.

SNOMED CT is a KOS based on OWL-EL and can thus be seen as a hierarchy of classes. All its RUs are *FormalRepresentationalUnit* as they all have axiomatic representations. SNOMED CT allows post-coordination, then also exhibits composite representation as level 1 RUs. SNOMED CT concepts can be seen as extensions of *AristotelianUniversals* or *TypeByIntention*, although in some cases such as 249820005 – "Absence of toe (finding)", a full logical definition, covering the intended meaning of this RU, is not given due to the lack of negation support in OWL-EL. For each RU, most terms are synonyms or near-synonyms with the fully specified names in each language being the preferred label. Textual composite representations are rare. Another particularity is that, in SNOMED CT, everything is named "concept", even those RUs that correspond to binary predicates and which are represented as OWL object properties.

HL7 hl7VS-appointmentReasonCodes, along with many other so-called value sets of the

[6] Most of it may be interpreted as **is_a**

HL7 standard is here presented as an example of a minimalist form of a KOS, consisting only of a *flat list* of RUs, here ROUTINE, WALKIN, CHECKUP, FOLLOWUP, and EMERGENCY. The labels correspond to the ID of each RU. All RUs in this KOS have free-text definitions. The RUs denote Aristotelian universals as they can all be instantiated by a particular appointment.

4. Discussion

Related work. KOSonto is the first, strictly ontology-based, attempt to lay a foundation for a principled ontological account of KOSs, in order to support interoperability and data integration in a domain characterized by the use of numerous KOSs with different structures, semantics, partly overlapping content, and diverging use cases. KOSonto is built under BFO and IAO. In IAO, referents are restricted to particulars, following BFO as an ontology of particulars, although BFO has never clearly committed to the representation of portions of reality beyond particulars. However, not all biomedical KOSs of interest are committed to ontological realism [23] – as claimed for OBO Foundry ontologies – and not all discourses in science and health have only particulars or classes of particulars as referents [24]. KOSonto addresses this by extending BFO beyond the *Continuant* / *Occurrent* bipartition, by introducing the common parent *Particular*, which is then juxtaposed to *Type* and *Class.* The consideration of types as "first-class citizens", besides particulars, can also be found in other upper-level approaches, e.g., Lowe's four-category ontology [25], as well as in the foundational ontologies GFO and UFO [26]. We have also granted this role to *Class* because it is a central element of KOSs and it is an implementation-specific construct in KOSs. Classes facilitate categorization and organization and can be seen as a specific manifestation or implementation of a type, not the type itself. Despite the focus on BFO and IAO in this work – because of the need to represent ICEs – we aim at the compliance of our approach with other ontological frameworks, including the Ontology of General Topology (OGT), and the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). KOSOnto deliberately does not use domain-based frameworks as support, such as the semiotic triad [27] or "Ontoterminology" [28], as it prioritizes practical relevance for KOS harmonization over attempts to achieve a painful reconciliation of the diverging philosophical views.

Main findings. KOSonto and its application to known KOSs highlight how KOSs of the most diverse types can be described: from small, non-hierarchical to mono- and multi-hierarchical systems, from informal to formal ones, and from reality-based to language-centered ones. Besides providing descriptions for the architectural constituents of KOSs, KOSonto particularly accounts for a typing of the different kinds of referents and their relevance for KOSs: types whose instances are ICEs, types whose instances are particulars, types that do not commit to either being hypothetical or real, non-empty classes of particulars or information entities, non-ontological relations such as **is_broader_than**. *Formal KOSs* can be described as KOSs that contain some *FormalRepresentationalUnit* and are expressed in a language based on logic, such as OWL, OBO syntax, or FOL. However, it is important to note that KOSonto has so far centered on referents that focus on *what kind* of entity is referenced, and not on the "targeted referents", i.e. *what particular entity proper* is meant by the use of an RU in a health record [29]. The changing nature and ambiguity raised by the "targeted referent" represent facets associated with the practical use of RUs, which our framework has not specifically addressed for now.

However, by clearly distinguishing different types of referents, it should be compatible with the ideas underlying the cited "Referent Tracking" paradigm. For example, in John Doe's record, the ICD-11 code "2D10.Z" – within an instance of the FHIR resource *Condition* – has a particular cancer in John Doe's thyroid gland as its referent, which is an instance of **Thyroid cancer**$_{type}$ (equally a member of the class *Thyroid cancer*, which is the extension of **Thyroid cancer**$_{type}$). In contrast, in another Jane Doe's record, the same ICD-11 code is present, but in a FHIR *Condition* instance where the slot **verificationStatus** is filled by *Refuted.* Here, the referent is not a particular entity (there is no cancer in Jane Doe's thyroid) but the type **Thyroid cancer**$_{type}$ itself, referred to in a negation statement (an ICE instance). This points to the potential of KOSonto in supporting the work on referent tracking. KOSonto also addresses the complex topic of identifiers, lexical features, and textual composite representations of RUs in KOSs. While many classical ontologies are not concerned with lexical features of RUs beyond unique IDs and human-readable labels (e.g., most Gene Ontology classes do not have synonyms), other KOSs (e.g., MeSH) have their focus on a high coverage of lexical units such as variants, synonyms, and entry terms. When describing KOSs, we consider the lexical properties such as labels, synonyms, and textual composite representations as orthogonal to the ontological aspects. Large parts of SNOMED CT are rich in ontology axioms and also in synonyms and entry terms. Textual composite representations are only present in a small part of SNOMED CT. On the other hand, there are small KOSs, such as the HL7 value set described above, which lack both ontological and lexical richness and are limited to unique labels and textual composite representations. This shows that an easy, mono-axial categorization of KOSs, e.g., in the sense that axiom-rich ontologies are at one end of the spectrum and lexicon-rich informal thesauri at the other, is not satisfactory. Further work will require to align KOSOnto with the descriptions of lexical features of KOSs, such as OntoLex-Lemon [30].

Limitations. We have intentionally left out all aspects of metadata (provenance, version, editorial notes, authors, etc.), as there are already initiatives such as the PROV Ontology[7] or Metadata vocabulary for Ontology Description and publication (MOD) [31]. We deliberately did not elaborate on the quality issues of KOSs either. Particularly, the mismatch between labeling and implicit meaning of KOSs in a particular scenario of use is a known issue [32], particularly due to numerous exclusion rules in classification systems such as ICD-10, so that labels can no longer be interpreted literally. There is also the problem of fuzzy and even misleading labeling of key concepts such as "Clinical finding" or "Qualifier value" in SNOMED CT [33]. Another quality issue is the misuse of formal languages such as OWL to express thesaurus-style content, driven for example by the popularity of the Protégé ontology editor, which seduces users into creating frame-like knowledge models without being aware of the far-reaching consequences of logical inference, cf. [34] for the NCI thesaurus. Similar issues would arise in implementing classification systems like ICD-10 in OWL [35]. A detailed analysis of KOS quality issues is currently not in the scope of KOSonto. On the other hand, our decision to use OWL for the description of KOSonto limited the expressivity of the ontology. This is a pragmatic compromise by the authors who recognize the fact that, despite the reasons mentioned above for using OWL, even OWL is not always well understood and implemented in its full expressivity.

[7] https://www.w3.org/TR/prov-o/

5. Conclusion and future work

By developing KOSonto, we responded to the need for a principled analysis and description of KOSs in the biomedical field. The modeling principles of KOSonto are of important significance, as they lay the foundation for a deeper understanding of how biomedical language and discourse, biomedical entities in reality, and representational artifacts are interconnected. Recognizing that KOSs are an extremely heterogeneous class of knowledge artifacts, built by different communities, for different purposes, on different knowledge organization traditions and using different architectures, it is difficult to foresee a convergent evolution in the near future. Therefore, the need for mapping between different KOSs becomes inevitable, demanding a common framework. The proposed ontology not only characterizes the representational artifacts themselves but also delves into their relationships with a wide range of referents, spanning from real entities to hypothetical and even fictional entities, all of which hold relevance within healthcare and life science discourse. Moving forward, the next crucial phase entails evaluating the suitability of the KOSonto framework for formally describing and facilitating mapping and harmonization activities across diverse KOSs. This evaluation will contribute to the ongoing search for common ground and improve the effectiveness of creating, using, and reusing mappings between heterogeneous KOSs.

6. Acknowledgements

We would like to express our deepest gratitude to the CRESP research center, for their support and provision of accommodations during the development of this article.

References

[1] S. Schulz, J.-M. Rodrigues, et al., Interface terminologies, reference terminologies and aggregation terminologies, Stud Health Technol Inform 245 (2017) 940–944.

[2] A. Isaac, E. Summers, SKOS simple knowledge organization system primer: W3C working group note 18, 2009. URL: https://www.w3.org/TR/skos-primer/.

[3] Medical Informatics Vocabulary (MIVoc). iTeh standards store, 1997. URL: https://standards.iteh.ai/catalog/standards/cen/beb23db9-36ca-4283-aaad-79ff90535f0f/env-12017-1997.

[4] F. Dhombres, J. Charlet, et al., Knowledge representation and management, it's time to integrate, Yearb Med Inform 26 (2017) 148–151.

[5] J. N. Nikiema, V. Jouhet, et al., Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts, J Biomed Inform 74 (2017) 46–58.

[6] A. T. McCray, S. J. Nelson, The representation of meaning in the UMLS, Methods Inf Med 34 (1995) 193–201.

[7] HeTOP, CISMeF, 1997. URL: https://www.hetop.eu/hetop/.

[8] N. F. Noy, N. H. Shah, et al., BioPortal: ontologies and integrated data resources at the click of a mouse, Nucleic Acids Res 37 (2009) W170–W173.

[9] L. Zheng, Z. He, et al., A review of auditing techniques for the UMLS, J Am Med Inform Assoc 27 (2020) 1625–1638.

[10] G. Bajaj, V. Nguyen, et al., Evaluating biomedical word embeddings for vocabulary alignment at scale in the UMLS Metathesaurus using siamese networks, in: Proc 3rd Workshop on Insights from Negative Results in NLP, 2022, pp. 82–87.

[11] J. N. Otte, J. Beverley, A. Ruttenberg, Basic Formal Ontology, Appl Ontol (2022) 1–27.

[12] B. Smith, W. Ceusters, Aboutness: towards foundations for the information artifact ontology, in: Proc of the 6th Intl Conf on Biomed Ontologies, 2015, pp. 1–5.

[13] E. M. Sanfilippo, Ontologies for information entities, Appl Ontol 16 (2021) 111–135.

[14] B. Smith, W. Kusnierczyk, et al., Towards a reference terminology for ontology research and development in the biomedical domain, in: CEUR Proc, volume 222, 2006, pp. 57–65.

[15] B. Glimm, I. Horrocks, al., HermiT: an OWL 2 reasoner, Journal of automated reasoning 53 (2014) 245–269.

[16] T. Sider, Ontological realism, Metametaphysics (2009) 384–423.

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

[18] A. D. Irvine, H. Deutsch, Russell's paradox (1995).

[19] R. Arp, B. Smith, et al., Building ontologies with BFO, MIT press, 2015.

[20] B. Smith, W. Ceusters, et al., Relations in biomedical ontologies, Gen Biol 6 (2005) 1–15.

[21] R. Hoehndorf, A. Oellrich, et al., Relations as patterns: bridging the gap between OBO and OWL, BMC Bioinformatics 11 (2010) 441.

[22] S. Schulz, J.-M. Rodrigues, et al., What's in a class? Lessons learnt from the ICD–SNOMED CT harmonisation, Stud Health Technol Inform 245 (2014) 1038–1042.

[23] D. Chalmers, Ontological anti-realism, Metametaphysics: New essays on the foundations of ontology (2009) 77–129.

[24] S. Schulz, M. Brochhausen, et al., Higgs bosons, mars missions, and unicorn delusions, in: Proc 2nd Intl Conf on Biomedical Ontologies. CEUR Proc, volume 833, 2011, pp. 183–189.

[25] E. J. Lowe, The four-category ontology, Clarendon Press, 2005.

[26] S. Borgo, A. Galton, et al., Foundational ontologies in action, Appl Ontol 17 (2022) 1–16.

[27] P. T. Raggatt, The dialogical self and thirdness: A semiotic approach to positioning using dialogical triads, Theory & Psychology 20 (2010) 400–419.

[28] C. Roche, Ontoterminology, in: Proc 8th LREC, 2012, pp. 2626–2630.

[29] W. Ceusters, The place of referent tracking in biomedical informatics, in: Terminology, Ontology and their Implementations, Springer, 2022, pp. 39–46.

[30] J. Bosque-Gil, J. Gracia, et al., The OntoLex Lemon lexicography module. Final community group report, 2019.

[31] B. Dutta, A. Toulet, et al., New generation metadata vocabulary for ontology description and publication, in: Metadata and Semantic Research, Springer, 2017, pp. 173–185.

[32] M. Kreuzthaler, M. Brochhausen, et al., Linguistic and ontological challenges of multiple domains contributing to transformed health ecosystems, Front Med 10 (2023) 1073313.

[33] S. Schulz, R. Cornet, et al., Consolidating SNOMED CT's ontological commitment, Appl Ontol 6 (2011) 1–11.

[34] S. Schulz, D. Schober, et al., The pitfalls of thesaurus ontologization–the case of the NCI thesaurus, in: AMIA Annu Symp Proc, 2010, pp. 727–731.

[35] A. Rector, S. Schulz, et al., On beyond Gruber:"Ontologies" in today's biomedical information systems and the limits of OWL, J Biomed Informatics 100 (2019) 100002.

Reuse and Enrichment for Building an Ontology for Obsessive-Compulsive Disorder

Areej Muhajab[1,*], Alia I. Abdelmoty[1] and Athanasios Hassoulas[2]

[1]*School of Computer Science & Informatics, Cardiff University, Wales, United Kingdom*
[2]*School of Medicine, Cardiff University, Wales, United Kingdom*

Abstract

Building ontologies for mental diseases and disorders facilitates effective communication and knowledge sharing between healthcare providers, researchers, and patients. General medical and specialized ontologies, such as the Mental Disease Ontology, are large repositories of concepts that require much effort to create and maintain. This paper proposes ontology reuse and automatic enrichment as means for designing and building an Obsessive-Compulsive Disorder (OCD) ontology. The methods are demonstrated by designing and building an ontology for the OCD. Ontology reuse is proposed through ontology alignment design patterns to allow for full, partial or nominal reuse. Enrichment is proposed through deep learning with a language representation model pre-trained on large-scale corpora of clinical notes and discharge summaries, as well as a text corpus from an OCD discussion forum. An ontology design pattern is proposed to encode the discovered related terms and their degree of similarity to the ontological concepts. The proposed approach allows for the seamless extension of the ontology by linking to other ontological resources or other learned vocabularies in the future. The OCD ontology is available online on Bioportal.

Keywords

Ontology, OCD, Mental health, Conceptual modeling, Ontology enrichment, Ontology reuse,

1. Introduction

Obsessive-Compulsive Disorder (OCD) is a frequently debilitating and often severe mental health disorder that affects approximately 2% of the population[1]. The Royal Collage of Psychiatrists (RCPSYCH[2]) report that approximately 1 in every 50 people suffer from OCD at some point in their lives, amounting to about 1 million people in the UK, affecting men and women equally. It is also noted [1] that people could spend a long time struggling with the disorder, often hiding their symptoms, before they get appropriate help, possibly attributed to the shame or stigma associated with having disturbing thoughts (e.g. ego-dystonic sexual or violent) and compulsive behaviours . Coding information in electronic health records (EHR) using standard medical terminologies, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [2] can facilitate the efficient recording and integration of patient notes, ultimately leading

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

muhajaba@cardiff.ac.uk (A. Muhajab); AbdelmotyAI@cardiff.ac.uk (A. I. Abdelmoty); HassoulasA2@cardiff.ac.uk (A. Hassoulas)
0009-0007-9943-9169 (A. Muhajab); 0000-0003-2031-4413 (A. I. Abdelmoty)

[1] https://bestpractice.bmj.com/info/
[2] https://www.rcpsych.ac.uk/mental-health/problems-disorders/obsessive-compulsive-disorder

to more effective healthcare management [3]. However, existing clinical terminologies, such as SNOMED CT, and ontologies, such as the Mental Disease Ontology (MDO)[3] are limited with respect to the following dimensions.

1. **Semantic Richness:** Existing ontologies are evolving. No specific ontology or (sub-ontology) exist that delineates OCD, its types and diagnosing symptoms; enough for example to distinguish it from other related disorders, such as *illness anxiety disorder* or *hoarding disorder*[4]. The creation and evaluation of such resources is a costly exercise that requires both domain and technology experts.
2. **Semantic Heterogeneity:** The meaning and provenance of the terms or concepts used in these resources are not usually included or described in sufficient detail to explain the basis of its association with a particular classification hierarchy. For example, in Disease ontology (DOID), OCD is described as a type of *Anxiety Disorder*, whereas this classification has been updated in the DSM-5 revision in 2013, where it is now classified as a type of *Obsessive-Compulsive and Related Disorders (OCRDs).* Also, different classification hierarchies for the same concept is used in different ontologies. For example, *Obsession* in the MDO is a type of *Pathological Mental Process*, whereas it is a type of *Behavioral Symptom* in the Medical Subject Headings ontology (MeSH)[5], and a type of *Content of Thought* in SNOMED-CT. Understanding and establishing the similarity of concepts across ontologies is a well-known research challenge.
3. **Structural Richness:** Most clinical terminological resources and ontologies are described primarily with subsumption (IS_A) relationships and presented as large class hierarchies of concepts. The uncontrolled use of IS_A relation to signify different types of relations (such as PART_OF, IS_INSTANCE_OF, IS_ASSOCIATED_With, etc.) has been noted, e.g. in SNOMED CT [4], leading to structural overload and possible incorrect subsumption relationships. Also, some modelling constructs such as the use of multiple inheritance, where a concept can have multiple parent types, can lead to complexity in reasoning with the information.

In this paper, we propose the development of an OCD ontology to address some of the issues noted above. The methodology for development follows established proposals in the literature. In particular, structural richness is addressed by making use of rich ontological modelling in the logical definition of concepts, whilst semantic heterogeneity is addressed by the reuse of existing resources directly in the ontology as well as creating explicit reference to related concepts in other resources. Semantic richness is addressed by complementing the definition of concepts in the ontology by the automatic discovery of related concepts from relevant resources using deep learning. Ontology design patterns are proposed to encapsulate the links to other ontologies and the discovered related concepts. The resulting ontology consists of 97 classes (expanded to 1047 classes from other ontologies), 17 object properties (relationships) and 5 data properties. Bio_ClinicalBERT[6] and a text corpus from an OCD discussion forum are used in the ontology enrichment task. This work contributes to the efforts of building biomedical

[3] http://purl.obolibrary.org/obo/MFOMD.owl

[4] https://www.rcpsych.ac.uk/mental-health/problems-disorders/obsessive-compulsive-disorder

[5] https://www.nlm.nih.gov/mesh/meshhome.html

[6] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

ontologies, by providing modelling solutions that allow for the integration, reuse and enrichment of the ontological resources. The developed OCD ontology is available on Bioportal[7]. The remainder of the paper is organized as follows. Some related works on biomedical ontology development, reuse and enrichment are reviewed in section 2. Section 3 describes the proposed OCD ontology and outlines details of its development process. Ontology design patterns for reuse and enrichment are presented in section 4 , followed by conclusions in section 6.

2. Related Work on Ontology Reuse and Enrichment

A brief overview is given here of efforts in the field of biomedical ontologies reuse and enrichment.

An overview of the list of prominent biomedical ontologies is given in [5]. El-Sappagh et al [6] reviews the limitation and complexity of large clinical terminology resources such as SNOMED-CT and proposes its transformation to an ontology by aligning and reusing concepts from the Ontology of General Medical Science (OGMS). Top-level concepts are first manually mapped to OGMS, and then used to compose more refined (pre-coordinated and post-coordinated) concepts. More recently, the integration of ontological resources has become the focus of attention. For example,the Mondo Disease ontology aims to harmonize disease definitions across the world [8] by integrating the classifications and relationships of commonly used disease ontologies into a single semantically coherent resource. It employs a Bayesian approach to ontology structure inference, combining deductive and probabilistic inference, and aims to provide equivalence mappings with precise annotations. Another example of creating a large ontology, demonstrating the complexity and scale of the required effort, is demonstrated in the development of CIDO: the community-based coronavirus infectious disease ontology [7]. The methodology for development of the ontology is based on following the OBO Foundry ontology development principles, and utilizing the eXtensible Ontology Development(XOD) strategy, which prescribes: ontology term reuse, semantic alignment, use of ontology design patterns for new term generation, and community effort. A largely manual effort is ongoing into the development and visual analysis and evaluation of the resulting ontology. The process of encoding logical definitions manually when developing ontologies is a challenging task [8]. Ontology Design Patterns (ODPs) are defined as reusable modeling solutions to frequently occurring ontology design problems and are proposed as a useful tool to address this challenge. The complexity of the ontologies and the need for checking their consistency was investigated in [9]. Using the Foundational Model of Anatomy ontology, they analyzed the musculoskeletal content and show the inconsistencies in the use of relations, lack of definitions of relations, and incomplete definition of the hierarchy. They suggested the definition and use of ODPs to address these issues. Recent approaches are proposing the integration of ontologies by transforming them into a unified knowledge graph, that can be homogeneously queried with SPARQL endpoints [10]. Representing the ontologies as an RDF resource including all its entailment (all consequences of its logical definition) can help in the process of checking the similarity and consistency of the integrated resources.

[7] https://bioportal.bioontology.org/ontologies/OCD

[8] https://mondo.monarchinitiative.org

Ontology enrichment is a term that has been used in the literature from two points of view: a) enriching the ontology itself; that is extending the ontology by supplementing its existing structure with related terms and metadata, and b) enrichment by ontology; where the ontology is used as a source for discovering related concepts in a particular domain for a particular purpose. A common example of the latter task is the Gene Ontology (GO) term enrichment; a technique for interpreting sets of genes by making use of the GO system of classification, in which genes are assigned to a set of predefined bins depending on their functional characteristics [11].

In this paper, we are concerned with the first view point; enriching the ontology itself. Few research works have addressed this problem by utilizing existing resources, or other generic resources. For example, [12] used UMLS as the source of discovering synonyms for concepts in the ontology. In [13], deep learning with a large corpus of PubMed review articles and veterinary clinical notes was used to discover related terms to some pre-defined terms related to medical conditions. They then use the results to populate the UMLS Semantic Types and Groups ontology[9], whilst relying on specialized ontologies to represent the relatedness (e.g. lexical and provenance) relationships and properties. Utilization of standardised methods of linking ontologies to lexicographic resources[10] is an important aspect of this work. The research area of using ontologies and machine learning is still novel. An overview of how semantic similarity measures and ontology embeddings may be incorporated with ML methods is given in [14]. Further work needs to be done on exploiting the ontology structure, possibly by ontology embedding, in the task of ontology enrichment.

Several methodologies have been proposed to guide the process of developing ontologies. The NeOn [15] methodology places emphasis on collaboration and distributed development. It encourages the modularization of ontology building process, where domain experts contribute to different modules, while ensuring the overall consistency of the ontology. Systematic Approach for Building Ontologies (SABiO) [16] is a related approach that suggests a more guided workflow for the ontology development process, where the design of reference domain and operational ontologies is suggested and the reuse of foundational ontologies and pattern-oriented reuse are encouraged. The Open Biological and Biomedical Ontologies (OBO) Foundry proposed a set of principles and guidelines [17] for the development of ontologies to promote interoperability and standardization in the life sciences domains. OBO ontologies need to adhere to common design patterns and share a foundational set of relations, thereby fostering seamless integration and facilitating collaboration across diverse biomedical domains. The the eXtensible ontology development (XOD) methodology [18] is designed to be flexible and adaptable, allowing ontology developers to extend and customize it according to the unique characteristics of the target domain. By embracing a modular approach, XOD promotes the reuse of existing ontological components, which minimizes duplication efforts and ensures consistency across the ontology. In this paper, we build upon SABiO and XOD to demonstrate the reuse and enrichment of the OCD ontology. We incorporate elements from foundational ontologies, domain-specific ontologies, and external sources to enhance the OCD ontology, thus showcasing the potential of leveraging existing resources to enrich and expand knowledge representation effectively.

[9] https://lhncbc.nlm.nih.gov/ii/tools/MetaMap/documentation/SemanticTypesAndGroups.html

[10] https://www.w3.org/2019/09/lexicog/

3. The OCD Ontology Development Process

An overview of the ontology development stages is shown in Figure 1. In this paper, we present an overview of ontology building process. The ontology evaluation tasks include, a) evaluating the logical consistency of the ontology; checked using the DL reasoners in Protégé , b) evaluating the capability of the ontology to answer a set of competency questions defined in the knowledge acquisition phase; checked by formulating the questions using the SPARQL query language, and c) expert evaluation to judge the quality and coverage of the ontology. Detailed account of these tasks are left to future publications.

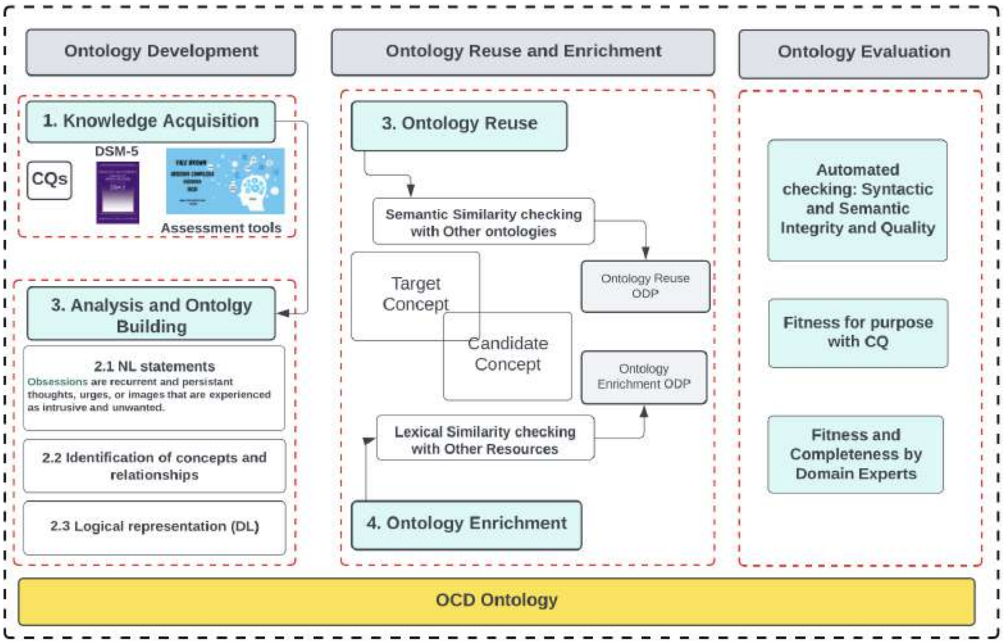

Ontology Development
Ontology Reuse and Enrichment
Ontology Evaluation
1. Knowledge Acquisition
CQs
DSM-5
Assessment tools
3. Analysis and Ontolgy Building
2.1 NL statements
Obsessions are recurrent and persistant thoughts, urges, or images that are experienced as intrusive and unwanted.
2.2 Identification of concepts and relationships
2.3 Logical representation (DL)
3. Ontology Reuse
Semantic Similarity checking with Other ontologies
Target Concept
Candidate Concept
Ontology Reuse ODP
Ontology Enrichment ODP
Lexical Similarity checking with Other Resources
4. Ontology Enrichment
Automated checking: Syntactic and Semantic Integrity and Quality
Fitness for purpose with CQ
Fitness and Completeness by Domain Experts
OCD Ontology

Figure 1: Ontology Development and Evaluation

Knowledge Acquisition: Two primary resources were used to acquire basic knowledge about the nature of the disorder and its diagnosis. These are the Diagnostic and Statistical Manual of Mental Disorder version 5 (DSM-5)[19] and OCD assessments tools, namely, the Yale-Brown Obsessive Compulsive Scale (Y-BOCS)[11] and the Obsessive-Compulsive Inventory (OCI) [12]. Statements describing OCD and its related concepts were manually extracted; 28 statements were extracted. Examples include, "OCD is characterized by obsession, compulsion, or both" and "the definition of obsession as an intrusive thought, image, or urge"; (both examples are from DSM-5). Y-BOCS, and other diagnostic tools, were particularly useful for identifying types of OCD and providing specific values of defined concepts. Eight types of *Obsession* and six types of *Compulsion* were identified. To further refine the definition of concepts that are not described in the primary resources, further relevant resources were employed. For example, the

[11] https://pcl.psychiatry.uw.edu/wp-content/uploads/2021/12/YBOCS.pdf
[12] https://www.div12.org/wp-content/uploads/2015/07/OCI.pdf

cognitive theory of OCD [20] emphasise that an intrusive thought transforms into an obsession when an additional meaning is attributed to it. A total of 35 statements were extracted (a full list can be found in GitHub repository "OCD-ontology". In this phase, we also compiled a set of competency questions (n=23) from diagnostic and other relevant resources to support the development and evaluation processes. Examples of the competency questions include: What patterns of activities do a person with aggressive thought has? and What is the type of obsession where avoidance behavior is a frequent occurrence?.

The Analysis Phase: Statements in the previous phase were used to identify relevant concepts and relationships. Statements for defining specific concepts were grouped and used to formulate a natural language definition that describe three questions: "what is the concept defined as?", "what are its types?", "what are its symptoms?". During this process, necessary and sufficient criteria for defining the concepts were identified. These refer to the properties (attributes and relationships) that must be present for an instance to be considered a member of the defined class. A natural language statement defining these properties is then formulated to allow for its transformation into logical statements in Description Logic (DL) and further definition in the ontology. An example of this process for the definition of the concept of *Obsession* is shown in Table 1. This approach allowed for a clear definition of the required concepts and for avoiding ambiguity and inconsistency in the representation by utilising the ontology reasoning tools over the defined ontology model. The analysis phase resulted in the definition of a set of 97 concepts and 22 types of relationships. This includes 17 object properties and 5 data properties.

Table 1
Example of the definition of the concept of *Obsession* in the analysis phase.

Statements relevant the concept of Obsession (1) Obsessions are recurrent and persistent thoughts, urges, or images that are experienced as intrusive and unwanted (DSM-5). (2) Individual with obsession attempts to ignore or suppress such thoughts, urges, or images, or to neutralize them with some other thoughts or actions (i.e., by performing a compulsion) (DSM-5) (3) Individual with OCD may experience over-importance of thoughts [20]. (4) There are 8 types of obsession (Y-BOCS).
Natural Language Definition In OCD, an Obsession can be any of: Intrusive Thought, Intrusive Image or Intrusive Impulse or Urge, that causes distress due to the added importance that the individual places on them. In OCD, Obsession is often accompanied by some Compulsions. **Description Logic Expression** $((\text{Obsession} \equiv \text{Intrusive thought} \sqcup \text{Intrusive image} \sqcup \text{Intrusive urge})$ $\sqcap (\exists\ \text{hasAssociatedAppraisal.ThoughtAppraisal}\ ))$
Related Identified Concepts *thought*, *intrusive thought*, *persistent thought*, *mental image*, *urge* and *thought appraisal.*

4. Ontology Design Patterns for Reuse and Enrichment

Reusing Existing Ontologies Related ontologies were identified by searching the NCBO BioPortal [13]; a comprehensive repository of biomedical ontologies. For every concept defined

[13] https://bioportal.bioontology.org/

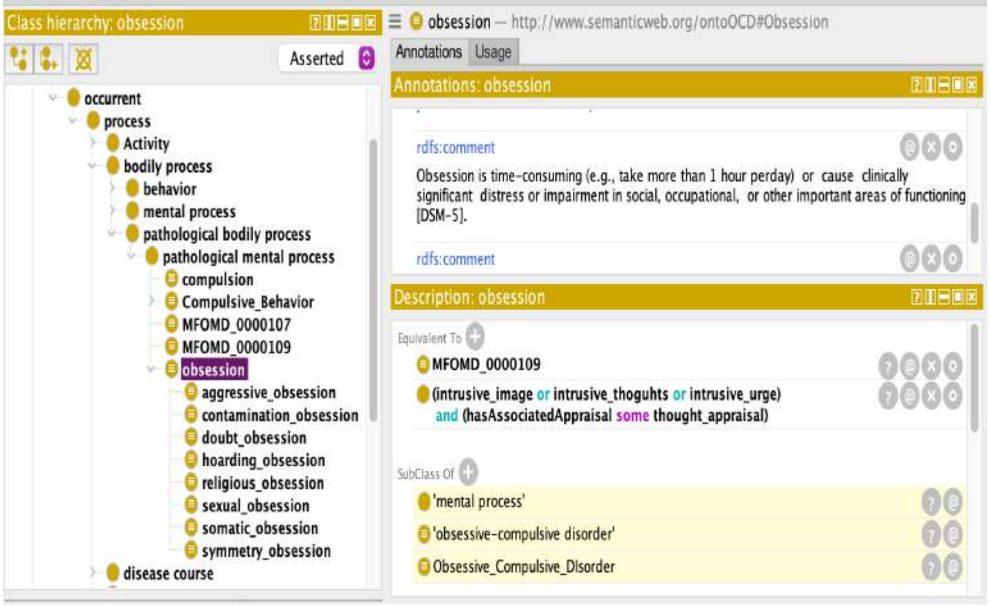
in the analysis phase, corresponding concepts were identified in the existing ontologies. The following heuristics were employed in the decision to reuse concepts.

1. The external concept is considered to be fully equivalent to the required OCD concept, if there is a complete overlap between the logical definitions of the two. In this case, the concept is imported directly to the ontology. When a concept is imported, all its related concepts, including its inheritance tree hierarchy, are also imported. For example, the *Activity* and *Symptom* classes are root classes in the Activity of Daily Living (ADL) ontology [21] and Symptom Ontology (SYMP) [14], respectively. Importing both classes in the OCD ontologies implies the use of their complete ontologies as well.
2. The external concept is considered to be partially equivalent to the required OCD concept, if its logical definition can be considered part of the definition of the required concept. For example, "OCD" is defined in the DOID ontology as a subclass of "Anxiety Disorder". No further definition is given in the DOID ontology. This definition is partially sufficient for our ontology and we need to further refine it. Hence, instead of importing the class and redefining it, we align our definition with the external ontology using the OWL:equivalentClass; an example is, ocd:OCD $\equiv$ DOID:OCD (where "ocd" and "DOID" are prefixes for the OCD and the Disease Ontology, respectively). This ontology alignment design pattern allows flexibility of ontology specification, whilst also reusing existing resources. There are 13 OCD concepts aligned as equivalent to external concepts. Figure 2 illustrates the refinement of the definition of the class "Obsession" in the OCD ontology. The definition of the "Obsession" in MDO is defined as: MDO:MFOMD_0000109 $\sqsubseteq$ MDO:Pathological mental process; which is defined as ($\sqsubseteq$ OGMS:Pathological bodily process $\sqcap$ ($\sqsubseteq$ $\exists$ manifestationOf.Mental disorder)) . This definition is reused in our ontology as follows: ocd:Obsession $\equiv$ MDO:MFOMD_0000109 (class obsession from MDO). The refinement of ocd:Obsession is presented as follows: ocd:Obsession $\equiv$ Intrusive thoughts $\sqcup$ Intrusive image $\sqcup$ Intrusive urge) $\sqcap$ ($\exists$ hasAssociatedAppraisal.ThoughtAppraisal); ocd:Obsession has 8 sub-classes.
3. The external concept is considered to be nominally similar to the required OCD concept, if there is some overlap between the logical definitions. In this case, we are unable to reuse the external class, but we maintain a link to it using the *Reuse Ontology Design Pattern*, as shown in figure 3.

The set of ontologies that were reused are as follow: Mental Disease Ontology (MDO)[15], Mental Functioning Ontology (MFO)[16], ADL, SYMP, Gene Ontology [17], SNOMED-CT, Experimental Factor Ontol- ogy (EFO), Gender, Sex,and Sexual Orientation(GSSO) ontology, Emotion ontology [18], and the Basic Formal Ontology (BFO)[19].

In the OCD ontology, classes that were not present in existing ontologies were created and mapped to classes in the Basic Formal Ontology (BFO) using the OWL:subClassOf relationship. The mapping process took into account the characteristics of each class in the BFO and

[14] https://obofoundry.org/ontology/symp.html
[15] http://purl.obolibrary.org/obo/MFOMD.owl
[16] http://purl.obolibrary.org/obo/MF.owl
[17] https://bioportal.bioontology.org/ontologies/GO
[18] http://purl.obolibrary.org/obo/MFOEM.owl
[19] http://purl.obolibrary.org/obo/bfo.owl



- occurrent
  - process
    - Activity
    - bodily process
      - behavior
      - mental process
      - pathological bodily process
        - pathological mental process
          - compulsion
          - Compulsive_Behavior
          - MFOMD_0000107
          - MFOMD_0000109
          - obsession
            - aggressive_obsession
            - contamination_obsession
            - doubt_obsession
            - hoarding_obsession
            - religious_obsession
            - sexual_obsession
            - somatic_obsession
            - symmetry_obsession
    - disease course

obsession — http://www.semanticweb.org/ontoOCD#Obsession

rdfs:comment

Obsession is time-consuming (e.g., take more than 1 hour perday) or cause clinically significant distress or impairment in social, occupational, or other important areas of functioning [DSM-5].

rdfs:comment

Equivalent To

MFOMD_0000109

(intrusive_image or intrusive_thoguhts or intrusive_urge) and (hasAssociatedAppraisal some thought_appraisal)

SubClass Of

'mental process'

'obsessive-compulsive disorder'

Obsessive_Compulsive_DIsorder

Figure 2: The representation of the class "Obsession" defined in protégé .

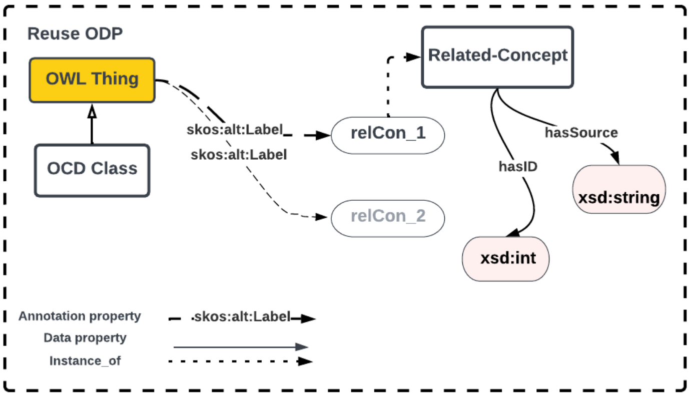
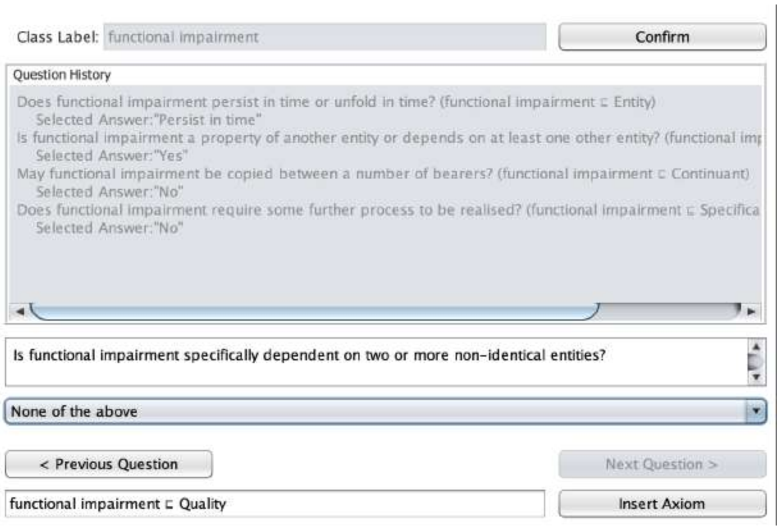


Figure 3: The Reuse Ontology Design Pattern.

determined the most suitable parent class for the OCD classes. A recent study by Emeruem et al. [22] proposed an automatic tool for mapping classes to the BFO. The aforementioned tools were used to guide the mapping of three classes in the OCD ontology: ocd:Severity Level, ocd:Assessment Criteria and ocd:Functional Impairment. Figure 4 illustrates the mapping of class "Functional Impairment" as ⊑ BFO:Quality.

Ontology Enrichment with Deep Learning

The objective here is to demonstrate how to automate the process of extracting related terms using deep learning from a given corpus for integration into the ontology. We first explore the



Class Label: functional impairment

Question History

Does functional impairment persist in time or unfold in time? (functional impairment ⊏ Entity)
    Selected Answer:"Persist in time"
Is functional impairment a property of another entity or depends on at least one other entity? (functional imp
    Selected Answer:"Yes"
May functional impairment be copied between a number of bearers? (functional impairment ⊏ Continuant)
    Selected Answer:"No"
Does functional impairment require some further process to be realised? (functional impairment ⊏ Specifica
    Selected Answer:"No"

Is functional impairment specifically dependent on two or more non-identical entities?

None of the above

functional impairment ⊏ Quality

Figure 4: The representation of class "ocd:Functional impairment" as ⊑ BFO:Quality based on the "Questions History".

efficacy of the BioClinical_BERT langugage model in identifying terms that are semantically similar to target terms in the ontology. BioClinical_BERT leverages contextual embeddings to capture the nuanced meaning of biomedical and clinical terminology. It is trained on a large corpus (2 million) of clinical notes. We then employ a word2vec model trained on an OCD-specific corpus to assess term similarity (cosine similarity). The word2vec model learns the distributional representation of words based on their co-occurrence patterns within the OCD corpus. *Candidate term extraction using BioClinical_BERT*: Target terms are mapped to their corresponding token IDs in the model and their hidden states are retrieved, representing the contextualized embeddings of the tokens. Similarity scores are then computed between the target term's contextualized embedding and the embeddings of other terms in the model. The top (n) terms with the highest similarity scores are then selected. A sample of candidates terms and similarity scores to target terms are listed in table 2. Notably, it was observed that certain target terms such as "compulsion", "intrusive" and "impairment" were not represented in the BioClinical_BERT model. *Candidate terms form the Word2vec model:* Here we utilized the word2vec model with the Continuous Bag of Words (CBOW) architecture to obtain a list of relevant terms for the same target terms in the ontology. Data for this study was collected from an OCD forum[20]. Python Selenium library was employed to gather a substantial dataset comprising 54,410 posts from the forum spanning the period from October 2000 to December 2021 . The experimental set-up; hyperparameter configuration, was as follows. The window size determines the range of neighbouring context words considered for each target word, and was set to 5. This means that words within a distance of 5 from the target word were taken into account. Additionally, a minimum count of 3 was defined, ensuring that words appearing at least three times in the dataset were included during training. Cosine similarity was then

[20] Online platform dedicated to OCD-related discussions and information exchange https://www.mentalhealthforum.net/forum/forums/obsessive-compulsive-disorder-ocd-forum.46/

computed between the target and candidate terms and the top(n) terms were selected. A sample of results is shown in table 3.

Table 2
Top similar terms to target terms from Bio_Clinnical BERT model

Target term	BioClinical_BERT (Top 6 terms)
obsession	(’obsessed’, 0.651), (’fascination’, 0.636), (’urges’, 0.588),(’irrational’, 0.573,), (’insistence’, 0.570) (’preoccupied’, 0.569)
urge	(’urging’, 0.573), (’urgency’, 0.535), (’encourages’,0.519), (’invite’, 0.496), (’desire’, 0.49), (’obsession’, 0.489)

Table 3
Top similar terms to target terms from Word2Vec model.

Target term	Word2Vec
obsession	(’theme’, 0.984), (’behaviour’, 0.983), (’compulsive’, 0.983), (’fear’, 0.979), (’trigger’, 0.974)(’habit’, 0.968)
compulsion	(’rumination’, 0.987), (’pattern’, 0.986), (’behaviour’, 0.984), (’ritual’, 0.979), (’activity’, 0.974), (’action’, 0.974).
urge	(’reflex’, 0.988), (’mindfully’, 0.984), (’opposite’, 0.979), (’reaction’, 0.976), (’deliberately’, 0.976), (’act’, 0.975)
intrusive	(’harm’, 0.9782), (’triggered’, 0.967), (’unwanted’, 0.965), (’invasive’, 0.965), (’horrific’, 0.965), (’disturbing’, 0.962)
impairment	(’indicative’, 0.962), (’distress’, 0.954), (’jealousy’, 0.944), (’inexplicable’, 0.940), (’deliberate’, 0.938), (’negativity’, 0.917)

An Enrichment ODP is proposed here, as in figure 5, to record the results. As shown in the figure, classes in the ontology can be associated with many alternative labels, whose properties, including, similarity score, method and date are also recorded. The pattern is a simple generic and flexible approach to associating multiple terms to the ontology. A more sophisticated approach to lexical representation of associated terms can also be envisaged, e.g. by employing the OntoLex ontology. This is the subject of ongoing work.

5. Conclusion

The process of defining an ontology is costly in terms of time and effort. Devising means of automating the process to complement the traditional approach will be of benefit to all stakeholders. This paper presents an approach for building an ontology for a specific mental disorder. The aim is to demonstrate how the traditional ontology building process can be complemented with a process of reusing existing ontology resources and enrichment with rich textual resources. The paper considers the building of an ontology for a specific mental disorder, as an example, but the approach proposed is generalisable to other use cases. A uniform approach is presented to enriching the ontology with ontological concepts and lexical terms using ontology design patterns. The degree of similarity of concepts in the ontologies guide the modelling process of the related concepts. Machine learning is used to discover similar terms to concepts

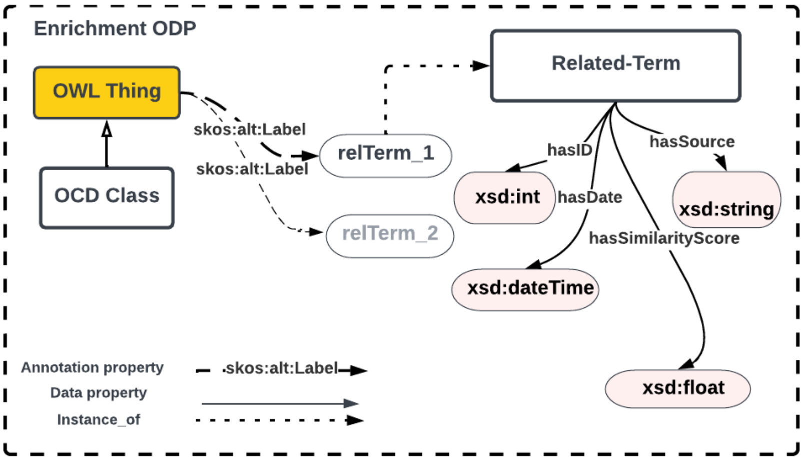


Figure 5: The Enrichment Ontology Design Pattern.

in the ontology using language models trained on large text corpus. The degree of similarity with the lexical terms are explicitly encoded. Results from the application of methods on two different corpora is presented. The paper outlines the approach and sets the way for further work on several fronts; refactoring the patterns to allow for richer modelling of lexical similarity, further refinement of the logical definition of the ontology based on expert evaluation. The detailed processes of building the ontology and its evaluation are the subject of future publications.

Acknowledgment

We would like to acknowledge the scholarship and support provided by Taif University.

References

[1] D. Veale, A. Roberts, Obsessive-compulsive disorder, Bmj 348 (2014) g2183.

[2] K. Donnelly, et al., Snomed-ct: The advanced terminology and coding system for ehealth, Studies in health technology and informatics 121 (2006) 279.

[3] M. Adnan, J. Warren, M. Orr, Ontology based semantic recommendations for discharge summary medication information for patients, in: 2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2010, pp. 456–461.

[4] O. Bodenreider, R. Cornet, D. J. Vreeman, Recent developments in clinical terminologies—snomed ct, loinc, and rxnorm, Yearbook of medical informatics 27 (2018) 129–139.

[5] J. D. Ferreira, D. C. Teixeira, C. Pesquita, Biomedical ontologies: Coverage, access and use, Systems Medicine (2020).

[6] S. El-Sappagh, F. Franda, F. Ali, K.-S. Kwak, Snomed ct standard ontology based on the ontology for general medical science, BMC medical informatics and decision making 18 (2018) 1–19.

[7] Y. He, H. Yu, A. Huffman, A. Y. Lin, D. A. Natale, J. Beverley, L. Zheng, Y. Perl, Z. Wang,

Y. Liu, et al., A comprehensive update on cido: the community-based coronavirus infectious disease ontology, Journal of Biomedical Semantics 13 (2022) 1–21.

[8] M. El Ghosh, F. Ghazouani, B. Birene, E. Akan, J. Charlet, F. Dhombres, Modeling logical definitions in biomedical ontologies by reusing ontology design patterns, in: ICBO'21: International Conference on Biomedical Ontologies, 2021, p. 20.

[9] M. D. Clarkson, L. T. Detwiler, K. M. Platt, S. Roggenkamp, Assessing the consistency of modeling in complex ontologies: A study of the musculoskeletal system of the foundational model of anatomy, Proceedings http://ceur-ws. org ISSN 1613 (2021) 0073.

[10] J. P. Balhoff, U. Bayindir, A. R. Caron, N. Matentzoglu, D. Osumi-Sutherland, C. J. Mungall, Ubergraph: integrating obo ontologies into a unified semantic graph, Proceedings http://ceur-ws. org ISSN 1613 (2022) 0073.

[11] A. Tomczak, J. M. Mortensen, R. Winnenburg, C. Liu, D. T. Alessi, V. Swamy, F. Vallania, S. Lofgren, W. Haynes, N. H. Shah, et al., Interpretation of biological experiments changes with evolution of the gene ontology and its annotations, Scientific reports 8 (2018) 5115.

[12] A. M. Rajput, H. Gurulingappa, Semi-automatic approach for ontology enrichment using umls, Procedia Computer Science 23 (2013) 78–83.

[13] M. Arguello-Casteleiro, R. Stevens, J. Des-Diz, C. Wroe, M. J. Fernandez-Prieto, N. Maroto, D. Maseda-Fernandez, G. Demetriou, S. Peters, P.-J. M. Noble, et al., Exploring semantic deep learning for building reliable and reusable one health knowledge from pubmed systematic reviews and veterinary clinical notes, Journal of biomedical semantics 10 (2019) 1–28.

[14] M. Kulmanov, F. Z. Smaili, X. Gao, R. Hoehndorf, Semantic similarity and machine learning with ontologies, Briefings in bioinformatics 22 (2021) bbaa199.

[15] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The neon methodology for ontology engineering, in: Ontology engineering in a networked world, Springer, 2012, pp. 9–34.

[16] R. de Almeida Falbo, Sabio: Systematic approach for building ontologies., Onto. Com/odise@ Fois 1301 (2014).

[17] 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.

[18] Y. He, Z. Xiang, J. Zheng, Y. Lin, J. A. Overton, E. Ong, The extensible ontology development (xod) principles and tool implementation to support ontology interoperability, Journal of biomedical semantics 9 (2018) 1–10.

[19] R. N. Kocsis, Book review: Diagnostic and statistical manual of mental disorders: (dsm-5), 2013.

[20] D. Julien, K. P. O'Connor, F. Aardema, Intrusive thoughts, obsessions, and appraisals in obsessive–compulsive disorder: A critical review, Clinical Psychology Review 27 (2007) 366–383.

[21] P. R. Woznowski, E. L. Tonkin, P. A. Flach, Activities of daily living ontology for ubiquitous systems: Development and evaluation, Sensors 18 (2018) 2361.

[22] C. Emeruem, C. M. Keet, Z. C. Khan, S. Wang, Bfo classifier: aligning domain ontologies to bfo, CEUR Workshop Proceedings (2022).

Version control for interdependent ontologies: Challenges and first propositions

Paul Fabry [1], Adrien Barton [2,1] and Jean-Francois Ethier [1*]

[1] *Groupe de Recherche Interdisciplinaire en Informatique de la Santé (GRIIS), Université de Sherbrooke, Quebec, Canada*

[2] *Institut de Recherche en Informatique de Toulouse (IRIT), CNRS, Université de Toulouse, France*

Abstract

The variety and quantity of health data have significantly increased due to healthcare improvements, making data interoperability an increasingly complex issue. Biomedical ontologies are a useful tool to support this interoperability. However, to provide an adequate coverage, ontological models can advantageously leverage a network of ontologies or part of ontologies, from various origins and strongly interdependent. While it fosters interoperability and minimizes duplication, this structure is very sensitive and any evolution of one of its parts requires more and more maintenance efforts. The challenge of tracking changes and assessing their effects is addressed in computer science through version control and while it has been applied to ontologies, applying this approach while importing parts of external ontologies and managing the potential impact of their evolution is a challenging task.

This article focuses on OWL ontologies and two questions regarding their versioning are addressed: what should be versioned in an ontology and how to determine its identity? The current processes rely on a versioning at the level of the ontology, which raises problems that are discussed. Versioning at a more fine-grained level of so-called “OWL components” could mitigate such problems. We identify two kinds of entities as potentially relevant to their identity: the associated cognitive representations and assertive statements; being able to track both separately would be an effective way to mitigate the aforementioned problems.

While further work is warranted to provide an operational definition of “OWL component”, an approach along the lines proposed here can support not only at a better change management through a cohesive, complete and fined-grained version control, but also a better import process while contributing to support ontology engineering methods.

Keywords

OWL ontologies, interoperability, version control

1. Introduction

Advances in modern healthcare have significantly increased the variety and quantity of health data generated. Ontologies are well positioned to serve as mediation models to support interoperability between diverse clinical data sources as they provide formal, source-independent representations and do not rely on specific data models that might be tied to specific technologies or formats. One such example is PARS3 [1], a distributed data access platform that is currently being deployed to support activities of the Health Data Research Network Canada [2] and to enable the implementation of a mortality prediction tool in hospitals in Quebec. Implementing these projects implies of series of tasks from ontology development, to pivot relational model generation, to mappings with data sources. While this approach has

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: paul.fabry@usherbrooke.ca (A. 1); adrien.barton@irit.fr (A. 2); jf.ethier@usherbrooke.ca (A. 3)
ORCID: 0000-0002-3336-2476 (A. 1); 0000-0001-5500-6539 (A. 2); 0000-0001-9408-0109 (A. 3)

been demonstrated and validated to enable the support of various tools like ReflexD [3], an audit-feedback tool for diabetes treatment in primary care, the next step is to support the evolution of users' needs. Science evolves, new requirements emerge, mistakes are identified, ideas are refined, etc. This will often imply changes in the knowledge representation generated at the inception of the project. This ontological evolution has a significant effect on the activities downstream of ontology development. It is critical to be able to identify what changes constitute a modification that can affect a user's understanding of what is represented or what should be instantiated in a table.

To follow on the questions above, if an ontology changes, how can one identify which of its parts might need to be reviewed for potential adjustments? If the answer is always "all of them", scalability is obviously at risk. What could be the consequences of this change on the applications which depend on it? In the context of PARS3, what are the impacts of a change in the ontology on the mappings between a data source and the pivot relational model? If the answer here again is to ask every source to review or redo every mapping between their assets and the pivot model, maintenance, scalability and as a result, acceptability and sustainability will be markedly reduced.

1.1. Background

This paper focuses on ontologies expressed in OWL2, as it is the most used Description-Logic language for biomedical ontologies. Good engineering principles have been proposed to mitigate the impact of ontology change. The Open Biomedical Ontologies Foundry [4] has proposed guidelines [5] including an emphasis on reusing the existing classes, a clearly defined scope for ontologies to ensure orthogonality of ontologies representing different domains and unique textual definitions following an Aristotelian form [6]. These principles were followed for the development of the ontologies supporting PARS3. A modular approach was chosen for the creation of the ontology model with several domain ontologies addressing specific topics such as drug prescriptions [7], laboratory results [8] or clinical forms [9]. All these ontologies are built upon the upper-level ontology Basic Formal Ontology (BFO) [10] and reuse classes and properties from reference ontologies in the biomedical domain like the Ontology of Biomedical Investigations (OBI) [11]. Finally, these ontologies are linked together via mid-level ontologies describing relevant parent categories pertaining to e.g. health procedures [12] or services [13].

Consequently, the clinical model structure consists of a network of ontologies or part of ontologies, which are either imported or created within our group, and are strongly interdependent. On the one hand it adds great flexibility and allows us to add representations of various domains as needed. In addition, ontology reuse enables us to leverage the domain expertise of other groups. On the other hand this structure is very sensitive and any change in its parts requires more and more maintenance efforts as the model grows and evolves, which may compromise its sustainability in the long run. Interestingly, the challenge of tracking change, evaluating its significance, and identifying the cascading effects is also an issue in computer science and is often addressed through version control.

Version control has already been discussed in the context of ontologies and is even part of the OBO Foundry principles [14] as well as FAIR [15]. However, in this context it consists in adding a version date in the metadata of the ontology. The change of the version date is at the discretion of the authors. Yet, different types of modifications in the ontology, such as the addition or removal of a class or the correction of an obvious typo in annotations such as a comment, can have very different impacts and it would be useful to differentiate between them. This is especially true since a version date refers to the whole ontology and does not in itself reveal which classes have been modified, which is all the more challenging in the case where a user imports only a few from a source ontology that may have several thousands.

The objective of this article is to lay the foundation for a versioning methodology that would allow the safe and efficient identification of what is impacted (or not) by a change in a context of highly interconnected ontologies. The core principles concerning both ontologies and versioning upon which this work is based and the issues at stake will be outlined, and some recommendations will be proposed and discussed. The resulting methodology and its implementation will be addressed in subsequent works.

2. Versioning challenges for ontologies

Version control as it pertains to computer science encompasses all the tools and methods for tracking and providing control over changes to a given project [16]. This control is performed at a granularity level deemed pertinent to manage the project most efficiently without creating an undue maintenance burden. To achieve this, artefacts of interest that are relevant to the project will be tracked individually. In many software projects, these artefacts are text files parts of the project's source code or documentation for example.

To ensure the tracking of an artefact, it is necessary to endorse a diachronic identity criterion on artefacts that allow us to identify some of them as evolutions of the same entity. An artefact like a computer file is identified (that is, uniquely denoted in a given context) by a resource identifier composed by the location in the file tree and the name of the file. This resource identifier provides a criterion of diachronic identity of the file it refers to, i.e. two files are considered as two states (versions) of the same informational artefact if they are identified by the same resource identifier. From a versioning perspective, even if the content of a file completely changes over time, the different versions will be considered as the same file at different states over time as long as they are identified by the same resource identifier. When trying to apply this methodology to ontologies, challenges emerge.

2.1. Ontology as the artefact of interest for versioning

One option could be to choose the ontology as the artefact of interest for the versioning. This calls for the characterization of what an ontology is. The word "ontology" can refer, depending on the context, to artefacts of various kinds, belonging to the fields of philosophy and computer science, which have as a common goal the representation of entities of the world or a particular domain of it, the categories to which they belong, their properties and the relations that hold between them.

In knowledge engineering, ontologies have been classically defined [17,18] as follows :

An ontology is a formal explicit specification of a shared conceptualization.

A definition grounded in a realist interpretation of the world is proposed by Smith et al. [10]:

Ontology = def. "A representational artefact, comprising a taxonomy as proper part, whose representations are intended to designate some combination of universals, defined classes, and certain relations between them."

The definition from [10] above assumes the existence of universals independently of the conceptualizations that we can make of the reality. In addition, an important element of Gruber and Guarino's definitions is the notion of shared conceptualization, that is an ontology is meant to be shared, usually between its authors and a community that is using it [19].

Many formalisms have been proposed for ontologies; among them, OWL is a widely used and recognized standard [20]. An OWL ontology is uniquely denoted with an internationalized resource identifier (IRI), which can be used for versioning purposes. For example, OBI is associated with the IRI: "http://purl.obolibrary.org/obo/obi.owl" that is specific to this ontology. While the ontology will undergo modifications over time, e.g. by adding and deleting classes or properties, it will be considered as the same ontology as long as its IRI remains the same. However, using the ontology as the versioning artefact may not be the best suited method.

Firstly, this approach considers an ontology as a monolithic artefact, one that is therefore used as a whole. This is a questionable assumption, especially in a context of interconnected ontologies where sometimes only parts of an ontology might be reused in another. Let's consider for example an ontology A that imports some part of an ontology B. If B changes, should a new version of A be declared even if not a single class declared directly in ontology A has been modified? If a new version of B is proposed, how to know if the classes imported in A are affected? If they are not, should this entail the creation of a new version of A?

Secondly, according to the W3C OWL2 reference documents [20], an ontology is defined as: "[…] a certain kind of computational artefact – i.e., something akin to a program, an XML schema, or a web page – generally presented as a document." While this allows a document-based approach to versioning, facilitated by tools such as Git [21], this might lead to counterintuitive consequences. For example, two classes with different namespaces located in the same computational artefact would be part of the same ontology according to the W3C definition above; however, in a different conception of ontology that would emphasize namespaces, they would belong to different ontologies.

Thirdly, the size of an ontology may represent a significant challenge for its versioning. The Gene Ontology [22] for example has more than 40 000 classes. If the modification of one of them leads to a new version of the whole ontology, this implies the tedious task for all its users of checking for possible changes in all the classes at each new version. Therefore, using the whole ontology as the most fine-grained artefact of interest for versioning is not an ideal solution.

2.2. Parts of ontology as the artefact of interest for versioning

An ontology proposes a representation of a domain by one or several authors. Such a representation takes place initially in the form of a collection of cognitive representations[2] of its author on the basis of which they create publicly available informational content entities (ICE) [23]. In the context of OWL2 ontologies, such ICEs will be called "OWL component" (OC). At least some of those OCs are associated with an identifier, such as an International Resource Identifier (IRI) that consists of a permanent location, a namespace and a unique string (for example: *"http://purl.obolibrary.org/obo/OBI_*0000011" is an IRI from OBI), which is of particular interest for versioning. Also at least[3] some of those OCs refer to a portion of reality that is part of the domain, in a way that can be understood by other users. These can be analyzed using a three level framework inspired by Ceusters and Smith [24]:

- Level L1: a portion of reality;
- Level L2: the cognitive representation of this portion of reality;
- Level L3: a publicly accessible ICE, the OWL component, that emanates[4] from the L2 cognitive representation and is also about the same portion of reality.

What is named in the literature as a 'class' or 'term' is an important type of such OCs. OCs can also include annotations such as definitions, or elucidations. Neuhaus [25] characterizes some of the annotations as "assertive annotations", i.e. "…the kind of annotations that are intended to assert a true proposition about the domain of the ontology." Neuhaus also regroup assertive annotations and logical axioms under the term of "assertive statements". A definition would specify in natural language a way to group certain particulars, while logical axioms can be automatically processed by a reasoner to classify the OWL components.

An OC can be associated with some logical axioms. For example, a class might be mentioned by some logical axioms.

For example, the OC with the IRI "http://purl.obolibrary.org/obo/OBI_0000011" (that has for label "planned process") is associated to the following assertive statements (among others):

[2] Cognitive representations are also concretizations of ICEs according to IAO [23].
[3] "at least" because it remains a complex open question whether all OCs are about a portion of reality (e.g. object properties).
[4] The precise nature of the relation of emanation is outside the scope of this work.

- Definition: “A process that realizes a plan which is the concretization of a plan specification.”
- *planned process* SubclassOf *process*

Two kinds of relations between a class and a portion of reality are proposed. Firstly, a relation of aboutness that is determined by the cognitive acts of its author: in the IAO framework, inspired by Chisholm [26], the “aboutness of those of our representations formulated in speech or writing […] is to be understood by reference to the cognitive acts with which they are or can in principle be associated” [23]. Secondly, a relation that is determined by the interpretation of the assertive statements by another competent reader who will make their own cognitive representation of a portion of reality on the basis of this interpretation. In this case, the assertive statements can be considered as describing that portion of reality (we will presume here that any competent external user will have the same interpretation). The goal when creating a class is that its assertive statements describe a portion of reality identical to the one pointed to by its aboutness.

2.3. Identity criteria for ontological components

One may propose two identity criteria of OC based on cognitive representations and assertive statements, depending on whether one relies on what we will call (inspired by theories in philosophy of language [27]) the “intentionalist” or “descriptivist” conceptions of the identity of an OC:

- R_1: In the intentionalist conception, the identity is determined by its author’s related cognitive representation.
- R_2: In the descriptivist conception, the identity is determined by its associated assertive statements.

Hence, in the case of R1, even though the assertive statements might change significantly, as long as the OC emanates from the same cognitive representation of the author, the identity of the OC does not change and so the IRI should remain the same. Meanwhile in R2, even though the author’s cognitive representation associated with an OC might change, as long as the assertive statements are not significantly modified[5], the identity of the OC does not change and so the IRI should remain the same.

The distinction between these conceptions is very rarely made explicit when creating and sharing an ontology, and this may negatively affect its evolution or its external reuse. The cognitive representation of the author (level 2) and the assertive statements are therefore two aspects that need to be managed appropriately to evaluate an OC as a potential artefact of interest for versioning, and as described above, two possible approaches, namely R1 and R2, need to be considered.

Let’s consider the following example about a hypothetical ontology about fruits, the Fruit Ontology (FO). At time t1 the author of this ontology adds an OWL component that refers to its own cognitive representation of the class of oranges (Citrus sinensis) in the reality. This OC (let’s call it “OC1”) is associated with the IRI: “FO_004” and is associated with the following assertive statements:

- Definition D1: “A citrus fruit which is the edible fruit of the orange tree and that is orange in color.”
- *orange* subClassOf *citrus fruit*

[5] We are talking about changes that might affect the understanding of the OC’s meaning, not merely a typo fix or a logically equivalent reformulation of a statement.

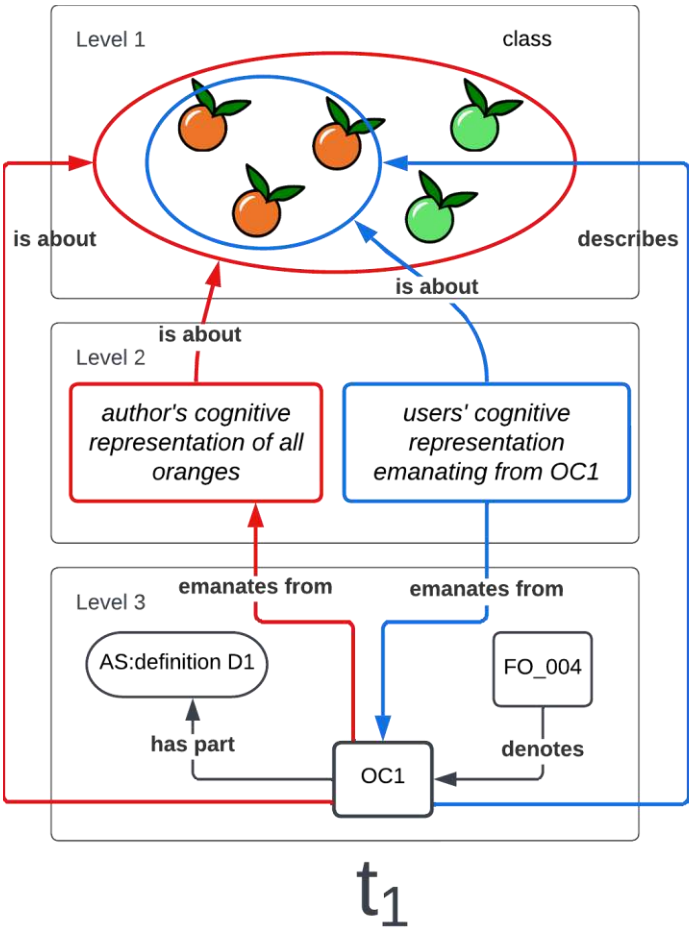


Figure 1: Relations between the OWL component (OC1) from the Fruit Ontology, the portion of reality it is about, and the associated author's and users' cognitive representations at time t1.

At time t2, the author is made aware that the definition D1 does not explicitly mention unripe oranges, which are green, and therefore that some of the ontology's users assume OC1 is about only ripe oranges (Figure 1).

The author is now faced with two options: either **A)** he considers that his initial cognitive representation is still the relevant one and so the assertive statements need to be updated accordingly, or **B)** he considers that what needs to be represented is instead ripe oranges and so he is now working with a new cognitive representation but as it turns out, the assertive statements derived from this new representation are similar to those initially derived.

Each of these options may have an impact on the identity of OC1 that must be evaluated according to the **R1** and **R2** conceptions discussed above, we thus have four distinct scenarios:

- Scenario **R1A**: The cognitive representation of the class of oranges is not modified between t1 and t2 and a new definition D2 ("A citrus fruit which is the edible fruit of the orange tree and that is orange in color when ripe, and green when unripe.") is proposed at t2 for clarification. In this scenario, OC1 is still conformant to the same cognitive representation, but one of the associated assertive statements has been modified. As the identity of OC1 is determined by the cognitive representation in R1 and this representation didn't change, what comes out of this scenario is a new version of OC1, which is thus associated with the same IRI "FO_004". (Figure 2).
- Scenario **R1B**: A new cognitive representation of interest (about ripe oranges) has been identified by the author of OC1. Therefore, a new OWL component (OC2) and thus a new IRI, is created by reusing assertive statements of OC1 (in this case the definition). (Figure 2).

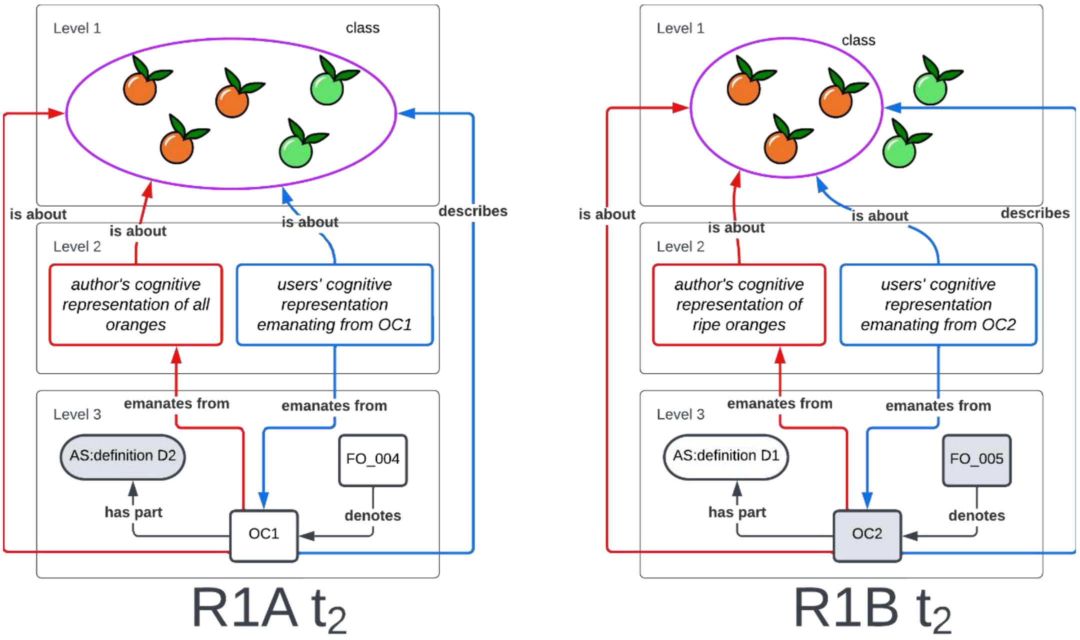


Figure 2: Scenarios R1A and R1B at t2. In R1A the cognitive representation of the class of oranges remains unchanged between t1 and t2 and a new definition is proposed at t2 for clarification. In R1B a new cognitive representation of interest has been identified by the author and therefore, a new OWL component (OC2) is created.

- Scenario **R2A**: R2A describes the same reality as R1A but endorses the R2 conception rather than the R1 conception. A new definition D2 has been formulated at t2, and consequently the assertive statements have changed. In this case, a new OC (OC3), and thus a new IRI, is created along the new assertive statements (Figure 3).
- Scenario **R2B**: Here also, R2B describes the same reality as R1B, but endorses the R2 conception rather than the R1 conception. The assertive statements have not been modified and thus the identity of OC1 has not changed even though OC1 is now related to another cognitive representation (Figure 3).

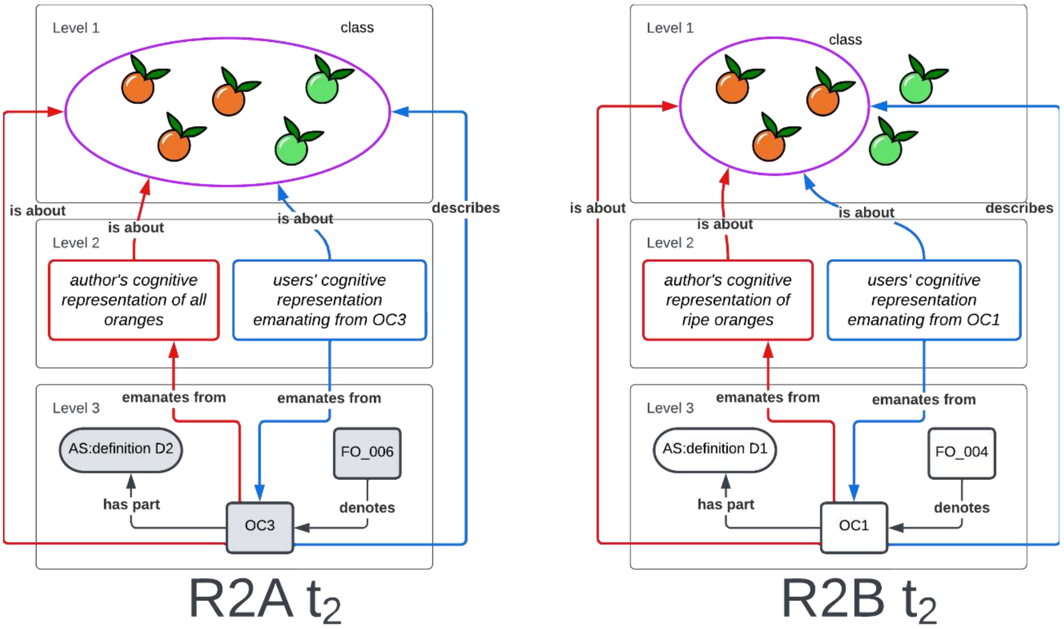


Figure 3: Scenarios R2A and R2B at t2. In R2A, the assertive statements have changed, thus a new OC and a new IRI are created along them. In R2B, the assertive statements have not changed. OC's identity has not changed even though it is associated with a new cognitive representation of the author.

Let's consider that OC1 is imported at time t1 in two fictive distinct ontologies:

- The Plant Parasites ontology (PPO), created by the author of OC1, imported it in the axiom: *medfly* subClassOf (is_parasite_of some *orange*). As a matter of fact, the author of PPO considered that FO_004 was about oranges at any stage of development.
- The Juice Ontology (JO) imported it in the axiom: *orange juice* subClassOf (is_made_of some *orange*). The authors of JO have initially imported "FO_004" as they understood the assertive statements as describing ripe oranges.

For these two ontologies, the previously mentioned scenarios will have various implication. For example, in the scenario R1A, as OC1 is now about oranges at any stage of development, JO needs to modify its axiom and include another OC while PPO doesn't need to. On the contrary, in the scenario R2B, as OC1 is now about ripe orange, JO could keep the axiom as is while PPO need to modify it. In context of a direct import, this will lead to incompatible OCs that might be difficult to identify, and the "live" change might affect users before the problem can be identified and fixed.

These scenarios are found in practice without being explicitly identified as such, which leads to many ambiguities when reusing ontologies. However, each conception has its advantages and disadvantages depending on whether one favors the author's cognitive representation importance in contributing to the coherence of the OC over the users' perspective which is based on their understanding of the assertive statements.

The examples above illustrate that OCs can be good candidates as artefacts of interest for versioning, provided that we are able to track both the cognitive representation of the author (via the operationalization of R1) and the assertive statements (via the operationalization of R2).

3. Discussion

Despite emerging tools that provide an invaluable help [28], managing the evolution of interconnected ontologies is a tedious and error-prone task. Versioning processes are a way forward to support the management of ontology evolution. In this work, two issues at the heart of this question are addressed.

3.1. Artefact of interest for ontology versioning

The first issue concerns the artefacts of interest, whose evolution needs to be tracked individually. As mentioned above, the current processes imply a versioning at the level of the ontology, but this presents some important problems, if only because it is not simple to outline the target of versioning in a context of interconnected modular ontologies where sometimes only parts of an ontology might be reused in another. Therefore, ontologies should no longer be considered as standalone and monolithic but rather as highly interconnected informational constructs, and a focus on a more fine-grained artefact is required, leading to the introduction of so-called “OWL components” as the versioned artefacts.

In the current work, we only discussed one kind of OC, namely classes. However, other components of an ontology such as object properties, data properties or even ontological instances are possible candidates as OCs of interest for versioning. Indeed, all these elements can be associated with assertive statements that can change over time. Defining OCs in the least ambiguous and most comprehensive way possible is of the utmost importance and will be the subject of subsequent work. In addition, not all parts of an ontology are of interest for versioning and therefore it will be important to determine whether certain OCs are not relevant for versioning.

It seems also relevant to consider what an ontology is in relation to OCs. On the one hand, ontologies are conceptually envisioned through high-level definitions such as the one previously mentioned [10,18]. On the other hand, ontologies as defined by W3C are described as computerized constructs and manipulated via files or namespaces. As it has been discussed in this work, the relationship between the two is not so clear cut and this calls for approaching OWL ontologies according to principles more in line with those of software engineering.

3.2. Identity of the artefact of interest

The next issue is the determination of the identity of the artefact of interest. Even tough the OCs relevant for versioning have only been characterized through their role for versioning, two possibly relevant aspects for their identity have been identified: their cognitive representations and their associated assertive statements (logical axioms and assertive annotations).

While the formal aspect of an ontology allows the use of reasoners and query tools to automatically process an ontology, an ontology is not limited to its formal elements. An ontology is also a way to share knowledge between communities of humans and for this goal, assertive annotations such as natural language definitions may have a role as significant as the logical axioms [19].

However, not all annotations are assertive and one needs to distinguish assertive annotations from annotations concerning the OC itself, such as its author or the date of creation. An annotation that will require additional exploration to arbitrate its inclusion or not as an assertive statement is the label of an OC. Given the uncertainty around its status, it is not included in the examples above. A natural language label in itself may be too ambiguous for playing an assertive role: experience shows that a given label could be associated with several distinct OC from distinct ontologies. However, labels are not randomly chosen either and the eventual import that the semantics of natural language labels should have into ontologies will need to be analyzed.

In an ideal world, one could envision a one-to-one relation between cognitive representations and assertive statements, as the cognitive representation of the author of an OC would be exactly translated in its assertive statements, which in turn would be exactly interpreted in a similar cognitive representation by the users that have only access to its assertive statements. However, OCs may retain the same IRI while their assertive statements undergo significant changes through time. Despite everyone's best effort, discrepancies exist and will likely remain and therefore the chosen versioning process must be able to handle these situations.

However, regardless of the quality of the assertive statements, an evolution of knowledge may imply a modification of the cognitive representation associated with an OC. Since OCs are developed in interlinked coherent groups (ontologies), the cognitive representation of the author is likely to play an important role in keeping this group coherent. Being able to track the cognitive representation of the author would allow the user community to know when something fundamental changed in the way the author structures the knowledge representation (e.g. because of a shift in science) versus when they attempt to improve the assertive statements to better match their cognitive representation.

To this end, being able to individually track the changes of both the cognitive representations and assertive statements would be a significant advantage. As it stands though only one IRI is associated with an OC and it is sometimes used to track the author's cognitive representation (R1) and sometimes the assertive statements (R2). However, it cannot do both in parallel. Associating two IRIs, one that tracks OC identity according to conception R1, and another one that tracks OC identity according to conception R2, would allow the tracking of both the cognitive representation and the assertive statements. Further work should elaborate the details of such a proposition.

4. Conclusion and Future Work

The work presented here aims at laying the foundation for a safe handling of changes to OWL components when they are reused. Being able to track changes in the cognitive representations of the authors or the assertive statements is the first step to further scalability and sustainability while diminishing the incentive to build "yet another model".

Other challenges lie ahead. It will likely be desirable to identify various types of changes without semantic implications including:

1. Changes to the computational representations e.g. changes of OWL syntax or in the order of elements;
2. Changes in non-assertive annotations e.g. additions of contact information annotations;
3. Changes in natural language assertive statements e.g. correcting a typographical error or replacing a word by a perfect synonym;
4. Changes in logical axioms producing a logically equivalent result e.g. replacing "OC_A OR OC_B" by "OC_B OR OC_A".

Automatically detecting changes of type 3 or 4 is not trivial in all cases.

Finally, when the changes include the cognitive representation of the author or assertive statements, it will be greatly beneficial to be able to have an approach to navigate through OCs of interest and identify those not impacted by the changes, leaving only a subset for manual review and adjustments as needed.

Overall, an approach based on the principles presented here will ensure not only a better change management through a cohesive, complete, and fined-grained version control, but also a better import process while contributing to supporting ontology engineering methods.

5. References

[1] PARS3 • Solutions • GRIIS. GRIIS n.d. https://griis.ca/en/solutions/pars3/ (accessed June 16, 2023).

[2] McGrail K, Diverty B, Lix L. Introducing Health Data Research Network Canada (HDRN Canada): A New Organization to Advance Canadian And International Population Data Science. IJPDS 2020;5. https://doi.org/10.23889/ijpds.v5i5.1493.

[3] ReflexD • Primary health care and follow-up for people with diabetes • Solutions • GRIIS. GRIIS n.d. https://griis.ca/en/solutions/reflexd/ (accessed June 16, 2023).

[4] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:1251. https://doi.org/10.1038/nbt1346.

[5] OBO Foundry - Principles overview n.d. http://obofoundry.org/principles/fp-000-summary.html (accessed June 16, 2023).

[6] Seppälä S, Ruttenberg A, Smith B. The Functions of Definitions in Ontologies. Formal Ontology in Information Systems 2016:37–50. https://doi.org/10.3233/978-1-61499-660-6-37.

[7] Ethier J-F, Goyer F, Fabry P, Barton A. The prescription of drug ontology 2.0 (PDRO): More than the sum of its parts. International Journal of Environmental Research and Public Health 2021;18. https://doi.org/10.3390/ijerph182212025.

[8] Barton A, Fabry P, Lavoie L, Ethier J-F. LABO: An ontology for laboratory test prescription and reporting. 2019 Joint Ontology Workshops Episode V: The Styrian Autumn of Ontology, JOWO 2019, vol. 2518, Graz; Austria: CEUR-WS.org; 2019.

[9] Fabry P, Barton A, Ethier J-F. QUESTO – An ontology for questionnaire. ICBO|ODLS 2020 International Conference on Biomedical Ontologies 2020, vol. 2807, Bolzano, Italy: CEUR-WS.org; 2020, p. B.1-12.

[10] Arp R, Smith B, Spear AD. Building ontologies with Basic Formal Ontology. Cambridge, Massachusetts: Massachusetts Institute of Technology; 2015.

[11] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. PLOS ONE 2016;11:e0154556. https://doi.org/10.1371/journal.pone.0154556.

[12] Fabry P, Goyer F, Barton A, Ethier J-F. An Ontological Analysis of Health Procedure Information. vol. 3073, 2021, p. 36–47.

[13] Fabry P, Goyer F, Barton A, Ethier J-F. An informational perspective on the ontology of services. vol. CEUR Workshop Proceedings, CEUR-WS.org; 2022, p. paper 5.

[14] OBO Foundry n.d. https://obofoundry.org/principles/fp-004-versioning.html (accessed March 30, 2023).

[15] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018. https://doi.org/10.1038/sdata.2016.18.

[16] Plaice J, Wadge WW. A New Approach to Version Control. IEEE Transactions on Software Engineering 1993;19:268–76. https://doi.org/10.1109/32.221137.

[17] Gruber TR. A translation approach to portable ontology specifications. Knowledge Acquisition 1993;5:199–220. https://doi.org/10.1006/knac.1993.1008.

[18] Guarino N, Oberle D, Staab S. What Is an Ontology? In: Staab S, Studer R, editors. Handbook on Ontologies, Berlin, Heidelberg: Springer; 2009, p. 1–17. https://doi.org/10.1007/978-3-540-92673-3_0.

[19] Neuhaus F, Hastings J. Ontology development is consensus creation, not (merely) representation. AO 2022;17:495–513. https://doi.org/10.3233/AO-220273.

[20] OWL 2 Web Ontology Language Primer (Second Edition) n.d. https://www.w3.org/TR/2012/REC-owl2-primer-20121211/ (accessed June 16, 2023).

[21] Chacon S, Straub B. Pro Git. Apress; 2014.

[22] Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, et al. The Gene Ontology resource: Enriching a GOld mine. Nucleic Acids Research 2021;49:D325–34. https://doi.org/10.1093/nar/gkaa1113.

[23] Ceusters W, Smith B. Aboutness: Towards Foundations for the Information Artifact Ontology. Proceedings of the Sixth International Conference on Biomedical Ontology (ICBO), CEUR vol. 1515; 2015, p. 1–5.

[24] Ceusters W, Smith B. Foundations for a realist ontology of mental disease. Journal of Biomedical Semantics 2010;1:10. https://doi.org/10.1186/2041-1480-1-10.

[25] Neuhaus F. What is an Ontology? 2018. https://doi.org/10.48550/arXiv.1810.09171.

[26] Chisholm RM. The primacy of the intentional. Synthese 1984;61:89–109. https://doi.org/10.1007/BF00485490.

[27] Michaelson E, Reimer M. Reference. The Stanford Encyclopedia of Philosophy 2022. https://plato.stanford.edu/archives/sum2022/entries/reference/ (accessed March 31, 2023).

[28] Matentzoglu N, Goutte-Gattat D, Tan SZK, Balhoff JP, Carbon S, Caron AR, et al. Ontology Development Kit: a toolkit for building, maintaining and standardizing biomedical ontologies. Database 2022;2022:baac087. https://doi.org/10.1093/database/baac087.

Ontology-based Integration of Consumer Data and EHR Systems to Fill Gaps in Social Determinants of Health Data

S. Clint Dowland [1], Melody L. Greer [1], Sudeepa Bhattacharyya [1,2], and Mathias Brochhausen [1]

[1] *University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA*
[2] *Arkansas State University, Jonesboro, Arkansas, USA*

Abstract

Social risk factors impact health outcomes. Understanding these risk factors requires increased collection and better maintenance of data on social determinants of health. Despite recent efforts to improve collection and organization of such data, there still are considerable hurdles to collecting these data in the clinical setting. To fill this gap, we propose extracting social determinants of health-relevant data from commercial consumer data as a source of additional individual-level, social risk factor-related data. We present early results of our efforts toward developing a social risk factor ontology and using an ontology-based approach to integrating commercial consumer data items with electronic health record data.

Keywords

social risk factor, social determinants of health, consumer data, ontologies

1. Introduction

It is widely recognized that social conditions influence health outcomes in many ways. These factors are called *social determinants of health* (SDOH). Due to the growing awareness of the importance of SDOH, efforts have been made to gather and organize data about them [1-6]. For example, there are SDOH-focused screening instruments such as the Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE); Kaiser Permanente's Structural Vulnerability Assessment Tool; and Epic's Healthy Planet module [7-9]. Additionally, there are codes for recording SDOH in coding systems such as SNOMED-CT and ICD-10 CM, and the Fast Healthcare Interoperability Resources (FHIR) data exchange standard includes SDOH condition categories [10].

However, studies have found that SDOH-relevant information is documented in clinical notes more often than through medical codes, that clinicians' lack training in gathering information about social risk factors as well as difficulties discussing sensitive subjects are obstacles to gathering SDOH data, and that some clinicians have concerns regarding the increase in administrative burden that comes with gathering social risk factor information [11, 12]. While there are area-level data that are relevant to SDOH, there are limits on the extent to which we can infer facts about an individual from these [13]. Greer, Zayas, and Bhattacharyya (2022) propose the use of commercial consumer data as an additional source of SDOH data as a solution to these problems [14].

In this paper we report early results on integrating commercial consumer data with electronic health record (EHR) data to address gaps in the availability of SDOH data for clinical and clinical research purposes. Our immediate goal is to create a pipeline from SDOH-relevant consumer data elements to EHR systems, thereby allowing health care providers to consider a more robust picture of a patient's social situation in a way that does not require gathering the data via additions to the workflow such as

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: clintdowland@gmail.com (S.C. Dowland); mlgreer@uams.edu (M.L. Greer); sbhattacharyya@astate.edu (S. Bhattacharyya); mbrochhausen@uams.edu (M. Brochhausen)
ORCID: 0000-0003-1909-9269 (S.C. Dowland); 0000-0002-1070-5907 (M.L. Greer); 0000-0003-3493-6029 (S. Bhattacharyya); 0000-0003-1834-3856 (M. Brochhausen)

using questionnaires. More broadly, we aim to identify SDOH-relevant consumer data and to enable the integration of SDOH data from disparate sources and of different types.

2. Background

To enable the pipeline from consumer data to EHR systems we aim to use an ontology-driven approach. By transforming relevant data items about a person into an ontology-enhanced form to represent phenomena the data items are about, as well as modeling FHIR-compatible medical codes in our ontology, we can make inferences about which of these codes characterize the person.

As the ontological basis for our project we are developing the Social Risk Factor Ontology (SRFON). SRFON is intended not only to represent types of entities and relations that are relevant to SDOH, but also to represent SDOH-related medical codes and link them to representations of the types of situations that they characterize. In these ways SRFON can enable the integration of SDOH-related data from disparate sources and inferences from data about an individual to medical codes that characterize that individual.

The above application of consumer data and SRFON requires ontologically modeling SDOH-relevant consumer data elements and medical codes, but does not require modelling EHR data because it is only intended to input information into EHR systems in the form of medical codes. This is advantageous for the purpose of providing information to health care providers in familiar formats, and saves time since fewer data elements need to be modeled and transformed. But we are also interested in the prospect of integrating SDOH-related data from consumer data sources with EHR data by ontologically modeling and transforming data of both types.

3. Materials and methods

Our methods are applied to two major components: a) developing an ontology covering SDOH, and b) identifying and selecting SDOH-relevant consumer data elements.

3.1. Developing SRFON

SDOH are not limited to direct influences on health, but instead include phenomena that influence health indirectly. In some cases the same phenomena have the potential to affect health through more than one pathway of influence, so that instead of only forming discrete causal pathways, SDOH can form interconnected webs of causes and effects. This web can include self-perpetuating cycles, such as when a person’s inadequate income is a barrier to transportation accessibility while the person’s lack of access to transportation is a barrier to employment opportunities that could lead to higher income. In order to develop an ontology to represent the relevant entities in the interconnected web of SDOH-related phenomena, we sought out academic literature that reports findings on how two or more SDOH are either correlated or causally related with one another or with effects on health. The starting point was a set of nineteen literature summaries from health.gov, each of which addresses a different SDOH area [15].

From these we extracted each assertion about pairs of social conditions and health effects that stand in a causal relation or are in some way correlated with each other. Many of the members of these pairs appeared in multiple assertions but denoted with different phrases, and so we identified such cases and made the phrasing uniform. In this initial list of terms for a number of interconnected phenomena, many terms did not name a single type of entity, but instead denoted a state of affairs involving multiple entities of certain types that are related in certain ways, and so each was analyzed in order to populate a list of terms for the relevant sorts of entities and relations. For example, household overcrowding is a matter of how several entities relate to one another, such as the members of some household and the rooms within their shared home.

Next, these terms were searched in three ontology repositories—Ontobee, BioPortal, and the Ontology Lookup Service—in order to find, when possible, preexisting ontology terms that represent the same types of entities [16-18]. Preference was given to terms from ontologies for which the Basic

Formal Ontology (BFO) is the upper ontology, as it is for SRFON [19]. Selected terms were imported, and new SRFON terms were created for types of entities or relations without matches. Terms in SRFON were arranged into a BFO-based hierarchy. Several SDOH-related clinical codes were included as well. These are represented as individual instances of *clinical code* and as members of their code sets.

3.2. Commercial consumer data and SDOH

Commercial consumer data include a wide range of types of information about an individual, the individual's household, and the area in which the individual resides. Commercial consumer data is gathered for the purpose of predicting a person's spending habits and it includes a vast amount of information about various aspects of the person's life.

In our project we use consumer data from a commercial database marketing company. Their database contains 6,260 distinct data elements that might each be populated with values for a given individual. We were provided with data dictionaries that include the value sets and written descriptions of each data element. We are manually reviewing these in order to find SDOH-related data elements. Additionally, to aid with finding and organizing relevant data elements, we have used keyword searches for SDOH-relevant terms, including but not limited to SRFON term labels and synonyms for them.

4. Results

4.1. SRFON

From the aforementioned literature summaries, we extracted 809 assertions about causal relations or correlations between pairs of phenomena including various social risk factors and health outcomes. Following the process of making the phrasing uniform and of replacing several terms that describe complex situations with corresponding collections of terms for the salient types of entities in those situations, there were 718 distinct class terms. After importing suitable terms from other ontologies—and in many cases importing superclasses of them as well—SRFON currently contains 677 new class terms and imports 255.

4.2. SDOH-relevant consumer data

While our review process is ongoing, we have thus far identified over 80 consumer data elements that are relevant to SDOH, either on their own or in combination. The consumer data include information about the employment status and education level of the person, each of which are important in relation to SDOH. In addition to the education level of the person the data is primarily about, there are also data elements about the education levels of up to four other individuals in the person's household. Additionally there are data elements about the occupation of the person and up to four other members of their household. Other information about the person's household can be derived from data elements about the total number of people in the household, the number of adults and the number of children in the household, whether there is a smoker in the household, and whether there is a single parent in the household. Relevant data about the person's home include the type of dwelling, whether the home has a source of heating or cooling, how many bedrooms and how many total rooms are in the home, and whether the home is owned or instead rented by the person. There are also data elements about the person's primary language and English proficiency, about the person's ethnicity at two levels of granularity, and how many vehicles are owned by members of the person's household. The consumer data also include area-level elements that are relevant to SDOH and specific to the area in which the person resides. These include for example seven cost of living indices at the county level, six of which concern the cost of specific types of products or services such as groceries, housing, and transportation.

The data elements described above are not an exhaustive list of SDOH-relevant consumer data elements, but suffice to reflect that information related to social risk factors can be derived at the individual level from commercial consumer data. Next, we take a closer look at some of these examples and how we ontologically represent what they are about.

4.3. Overcrowding

Household overcrowding occurs when too many people live together in the same residence, and it is a social risk factor [20-21]. The consumer data set we are utilizing does not contain any data elements that are explicitly about overcrowding nor any single value that indicates overcrowding on its own. However it includes data items about the person's household and residence that are relevant to measuring overcrowding.

Ways of measuring overcrowding tend to take as inputs both some measure of the household size and some measure of the home's capacity to house them [22]. For example, one standard that has been used is whether there are more than two persons per bedroom in the residence. In Figure 1, we represent a scenario in which some person P1's household consists of six members living in a residence with two bedrooms. In this figure, white nodes represent values of data items; blue nodes represent individual entities, including data items whose values are derived from consumer data; and green nodes represent individual entities whose values or relations to the other entities are inferred.

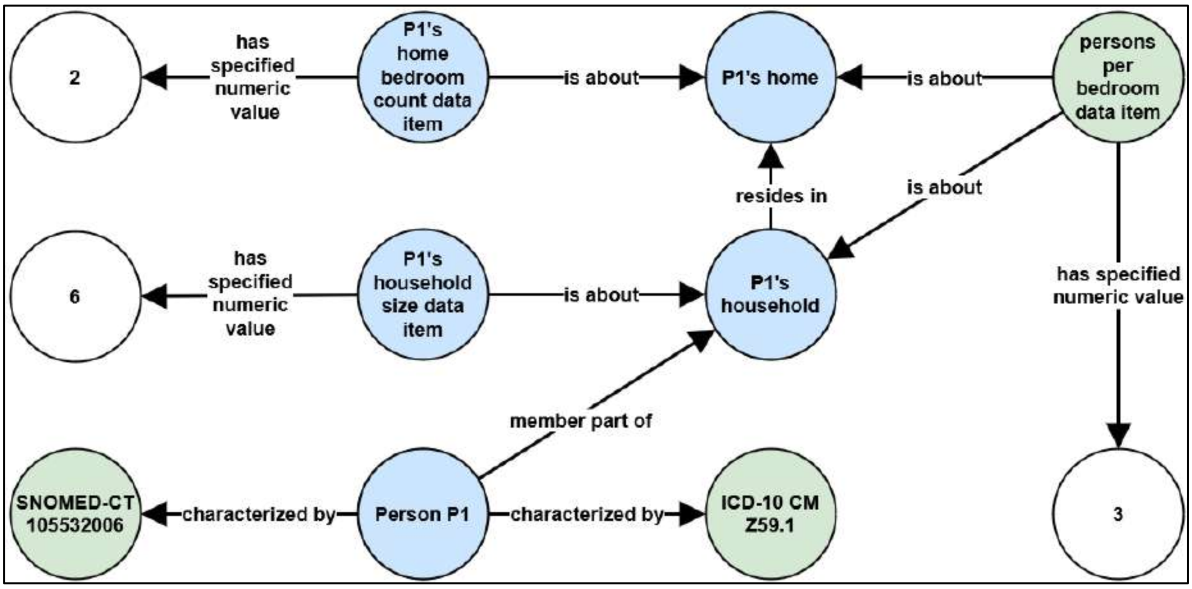

2
has specified numeric value
P1's home bedroom count data item
is about
P1's home
is about
persons per bedroom data item
resides in
is about
6
has specified numeric value
P1's household size data item
is about
P1's household
has specified numeric value
member part of
SNOMED-CT 105532006
characterized by
Person P1
characterized by
ICD-10 CM Z59.1
3

Figure 1: Person P1's household overcrowding

If we implement the aforementioned standard for assessing overcrowding, then from the combination of that standard and the inferred three persons per bedroom we can further infer that P1's household is experiencing overcrowding and thus that P1 is characterized by appropriate medical codes. For example, within the value set for the FHIR SDOH condition category 'inadequate-housing' are codes from both SNOMED-CT and ICD-10 CM. The SNOMED-CT codes include one that is specifically about overcrowding: 105532006, "Overcrowded in house." ICD-10 CM includes another that is applicable: Z59.1, "Inadequate housing." In Figure 1 we represent P1 as standing in the *characterized by* relation to each of these codes.

In addition to bedroom count and number of household members, the commercial consumer data also include the number of adults in the household and the number of children in the household, as well as the residence's number of rooms in general and its square footage. These are relevant for measuring overcrowding because, in addition to overcrowding standards that use the number of persons per bedroom, there are others that make use of the number of persons per room or the number of square feet per person, and some require a distinction between adult and child members of the household [22-23]. The consumer data is thus a potential source of information relevant to a number of ways that overcrowding has been measured. One advantage of this is that when the data required for one measure is not available for an individual, it might be possible to use a different measure. Another is the ability to compare and contrast different measures of overcrowding, for example by examining how often they evaluate the same households as overcrowded, or analyzing cases in which they evaluate the same households differently in order to investigate how other variables correlate with overcrowding as measured in different ways.

4.4. Language use and health care

Language barriers and limited English proficiency (LEP) can be detrimental to health in a number of ways. For example they can be obstacles to understanding health-related information from public sources [24]. Furthermore, language barriers between patients and providers are associated with lower quality of health care [25-26]. They can do so by inhibiting the patient's abilities to understand the provider's questions and to clearly convey problems and concerns to the provider, thus inhibiting the provider's ability to reach an accurate diagnosis. Additionally, language barriers can make it difficult for the patient to properly understand the provider's instructions once a diagnosis is made.

A terminological clarification will allow us to clearly distinguish two important concepts related to language use. By "primary language" we mean the language that a person is most adept at, or is most comfortable with, using. This is often the person's first language. In contrast, we use "preferred language" for the language a person selects to use in a given situation, for example during a health care encounter. These are often but not always the same, for preferred language can vary from situation to situation even while primary language stays the same. For example, a bilingual Spanish and English speaker might select English as their preferred language at a hospital in the U.S., while preferring Spanish when visiting a hospital in Mexico, so as to increase their chance of successful communication in each setting.

A preferred language field is found in EHR systems that meet Federal guidelines for stage 1 Meaningful Use Requirements [27]. But to know whether the preferred language is the patient's primary language and whether the choice of language might be cause for concern about a language barrier, we need more information. The commercial consumer data we are using can help with this because they include data about the person's primary language and about the person's ability to speak English. In Figure 2 we depict a scenario in which consumer data reflect that some person P2's primary language is Spanish and that P2 has LEP, while EHR data indicates P2 selected English as the preferred language for some particular health care encounter. The representation of this scenario in Figure 2 is based in part upon the way that languages, linguistic competences, and primary and preferred language data are represented in the Ontology of Medically Related Social Entities (OMRSE) [28].

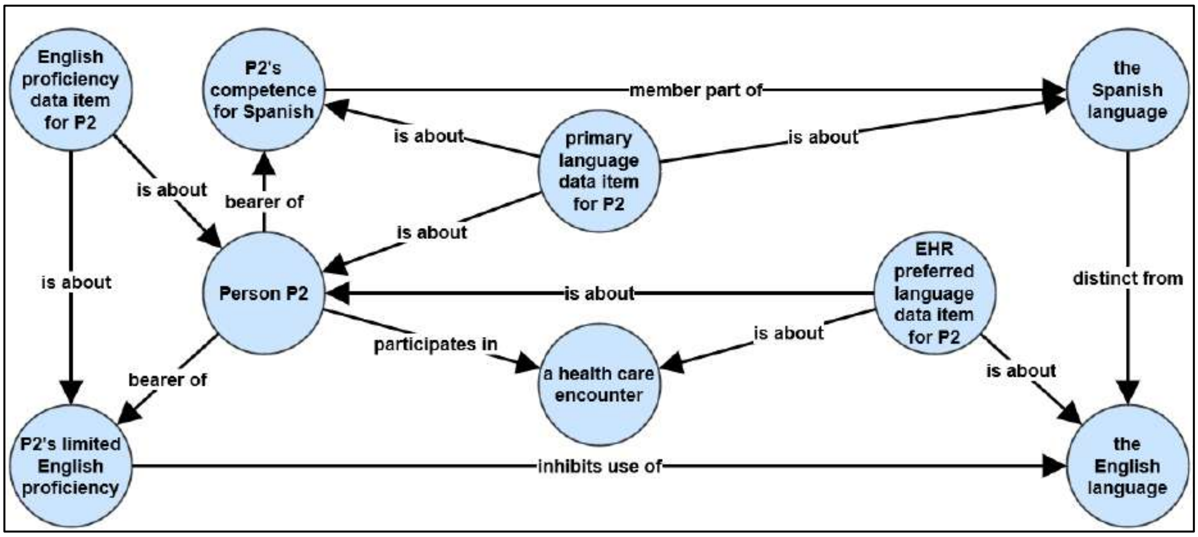


Figure 2: Data about P2's language use and capabilities

We can see that during the health care encounter, P2 was at risk for facing language-related barriers to the benefits of health care. Of course, not everyone whose primary language differs from their preferred language in a given encounter will face a language barrier during that encounter, since a person can be highly proficient in multiple languages. But having LEP and a primary language other than English indicates P2 faces a potentially detrimental language barrier when communicating with health care providers in English.

5. Conclusion

We have described a number of consumer data elements that are relevant to SDOH, and in more detail have discussed two sets of examples, how we represent what the data are about, and examples of what can be inferred from them. SRFON can aid in integrating such data with EHR or other SDOH-related data, as well as with enabling inferences from consumer data items to medical codes that characterize the person.

We will continue developing SRFON as well as identifying and ontologically modeling SDOH-related consumer data elements. Other future work includes using SRFON as the base for an additional ontological representation of SDOH-related correlations and causal relations that are reported as findings in academic literature—starting with those that informed the initial development of SRFON—thereby integrating findings from a number of sources. One possible use of this is to aid in identifying potential problem areas for individuals. For example, several variables in a person's life might each be of types that can cause or otherwise increase the risk of the same type of problem. Future work also includes looking into the potential utility of integrating individual-level consumer data with relevant area-level data from additional sources, such as from the US Census Bureau.

6. Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Award No. OIA-1946391. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

[1] Blackman, P. H. (1994). Actual causes of death in the United States. *Jama*, *271*(9), 659-660.

[2] McGinnis, J. M., Williams-Russo, P., & Knickman, J. R. (2002). The case for more active policy attention to health promotion. *Health affairs*, *21*(2), 78-93.

[3] Wilensky, G. R., & Satcher, D. (2009). Don't forget about the social determinants of health. *Health Affairs*, *28*(2), w194-w198.

[4] Braveman, P., & Gottlieb, L. (2014). The social determinants of health: it's time to consider the causes of the causes. *Public health reports*, *129*(1_suppl2), 19-31.

[5] Gold, R., Cottrell, E., Bunce, A., Middendorf, M., Hollombe, C., Cowburn, S., ... & Melgar, G. (2017). Developing electronic health record (EHR) strategies related to health center patients' social determinants of health. *The Journal of the American Board of Family Medicine*, *30*(4), 428-447.

[6] LaForge, K., Gold, R., Cottrell, E., Bunce, A. E., Proser, M., Hollombe, C., ... & Clark, K. D. (2018). How 6 organizations developed tools and processes for social determinants of health screening in primary care: an overview. *The Journal of ambulatory care management*, *41*(1), 2.

[7] PRAPARE. Who We Are. https://prapare.org/who-we-are/.

[8] Bourgois, P., Holmes, S. M., Sue, K., & Quesada, J. (2017). Structural vulnerability: operationalizing the concept to address health disparities in clinical care. *Academic medicine: journal of the Association of American Medical Colleges*, *92*(3), 299.

[9] OCHIN. Building the Foundation for Population Health at OCHIN. https://ochin.org/blog/population-health-at-ochin.

[10] HL7 International – Patient Care WG. (2023, June 02). SDOH Clinical Care: 14.6.1 Resource Profile: SDOHCC Condition.

[11] Guo, Y., Chen, Z., Xu, K., George, T. J., Wu, Y., Hogan, W., ... & Bian, J. (2020). International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. *Medicine*, *99*(52).

[12] Tong, S. T., Liaw, W. R., Kashiri, P. L., Pecsok, J., Rozman, J., Bazemore, A. W., et al. (2018). Clinician experiences with screening for social needs in primary care. J. Am. Board Fam. Med. 31, 351–363.

[13] Greer, M. L., Garza, M. Y., Sample, S., & Bhattacharyya, S. (2023). Social Determinants of Health Data Quality at Different Levels of Geographic Detail. *medRxiv*, 2023-02.

[14] Greer, M. L., Zayas, C. E., & Bhattacharyya, S. (2022). Repeatable enhancement of healthcare data with social determinants of health. *Frontiers in big Data*, *5*.

[15] US Department of Health and Human Services, & Office of Disease Prevention and Health Promotion. Social Determinants of Health Literature Summaries - Healthy People 2030. https://health.gov/healthypeople/priority-areas/social-determinants-health/literature-summaries.

[16] Xiang, Z., Mungall, C., Ruttenberg, A., & He, Y. (2011, July). Ontobee: A linked data server and browser for ontology terms. In *ICBO*.

[17] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., ... & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, *37*(suppl_2), W170-W173.

[18] Côté, R. G., Jones, P., Apweiler, R., & Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC bioinformatics*, *7*(1), 1-7.

[19] Arp, R., Smith, B., & Spear, A. D. (2015). *Building ontologies with Basic Formal Ontology*. MIT Press.

[20] Lepore, S. J., Evans, G. W., & Palsane, M. N. (1991). Social hassles and psychological health in the context of chronic crowding. Journal of Health and Social Behavior, *32*(4), 357–367.

[21] Cardoso, M. R. A., Cousens, S. N., de Góes Siqueira, L. F., Alves, F. M., & D'Angelo, L. A. V. (2004). Crowding: Risk factor or protective factor for lower respiratory disease in young children?. BMC Public Health, *4*(1), 1–8.

[22] Blake, K. S., Kellerson, R. L., & Simic, A. (2007). Measuring overcrowding in housing.

[23] OECD. (2023). Housing overcrowding (indicator). doi: 10.1787/96953cb4-en.

[24] Greer, M. L., Sample, S., Jensen, H. K., McBain, S., Lipschitz, R., & Sexton, K. W. (2021). COVID-19 is connected with lower health literacy in rural areas. *Studies in health technology and informatics*, *281*, 804.

[25] Espinoza, J., & Derrington, S. (2021). How Should Clinicians Respond to Language Barriers That Exacerbate Health Inequity? *AMA Journal of Ethics*, *23*(2), 109-116.

[26] Flores, G. (2005). The impact of medical interpreter services on the quality of health care: A systematic review. *Medical Care Research and Review, 62*(3), 255–299.

[27] Centers for Medicare & Medicaid Services (CMS), "Eligible Hospital and Critical Access Hospital Meaningful Use Core Measures Measure 6 of 11, Stage 1, Record Demographics."

[28] Dowland, S. C., Diller, M. A., Landgrebe, J., Smith, B., & Hogan, W. R. (forthcoming). Ontology of Language, with Applications to Demographic Data. *Applied Ontology*.

Beyond the Goods-Services Continuum

Peter Koch [1], Barry Smith [2],

[1] *Villanova University, 800 Lancaster Avenue, SAC 166, Villanova, Pennsylvania, United States*
[2] *SUNY @ Buffalo, 105 Park Hall, Buffalo, NY 14260, United State*

Abstract

Governments standardly deploy a distinction between goods and services in assessing economic health and tracking national income statistics, of which medical goods and services carry significant importance. In what follows we draw on Basic Formal Ontology (BFO) to introduce a third kind of entity called *patterns*, which help capture the various ways in which goods and services are intertwined and help also to show how many services generate a new kind of non-goods-related products. Patterns are an overlooked yet essential features of many economic sectors including medicine. Studying patterns offers new insights into various components of economic analysis, including outcomes-oriented evaluations of medical services and the value of human capital in the medical sphere.

Keywords

Goods, Services, Medical Care, Economics, Basic Formal Ontology (BFO)

1. Introduction: Goods, Service, and the Goods-Services Continuum

The income statistics created by different governments standardly deploy a distinction between goods and services in assessing the economic state of a nation. Gross Domestic Product, for example, is described by the United Nations as "a basic economic growth indicator [that] reflects changes in total production of goods and services." [1] On the one hand, *goods* are generally understood as alienable, often material things produced in an economic system, such as bottles of milk, songs, or combs – things that can be sold, lent, consumed, gifted, and so forth. On the other hand, *service*s are typically understood as tasks performed for the benefit of a consumer and are marked by the fact that production and consumption occur simultaneously [2,3]. When a person goes to a comedy show, for example, that person simultaneously consumes what the comedian produces (jokes) in exchange for money. Drawing upon the terminology codified in the Basic Formal Ontology (BFO),[2] goods are independent continuants, which preserve their identify over time; services are occurrents, which unfold through time and have successive temporal parts, such as when a mechanic services a car by draining the old oil, changing the oil filter, adding new oil, and so on [4]. We note here that goods are often used in the performance of a service – we shall see more generally in what follows that goods and services overlap.

In coming to a clear understanding of the ontology of goods and services, it is useful to differentiate between pure goods, which are those goods that are unaccompanied by a service, and pure services, which lack any accompanying good. Examples of the former are books and bracelets; examples of the latter are financial consultation services and tours of historical monuments. Between these two extremes are many items of economic significance that fall along a continuum between goods and services and in which goods and services are in various ways intertwined. For example, when you purchase a car,

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: peter.koch@villanova.edu (A. 1); smith@buffalo.edu (A. 2);
ORCID: 0000-0002-7923-0436 (A. 1); 0000-0003-1384-116X (A. 2);

[2] Basic Formal Ontology is a top-level realist ontology that conforms to the standards specified in ISO/IEC 21838-1. As an established top-level ontology, BFO provides a standardized framework for effectively acquiring, analyzing, and integrating information from multiple disparate domains. BFO has been adopted in over 500 ontology-based projects in domains such as biomedical sciences, genomics, and military research.

the act of purchasing often involves a salesperson who assists in selecting among different cars and different features of cars, such as color, performance specifications, and so forth. The salesperson may also give you an overview of how to operate features of the car and of how you can best pay for it. These are what we can call *supportive services*. A restaurant server will provide supportive services of taking your order, running the food to your table, and so on. The purchase of a car and the experience of a restaurant meal are composed in different ways of bundles of goods and services; they each fall along the goods-services continuum.

2. Beyond Goods and Services: Patterns as Products

While the distinction between goods and services and the accompanying idea of a continuum of cases which fall between them is useful and frequently invoked, we contend that this framework is in an important respect faulty. Consider the product provided by a masseuse. Typically, this product (a massage) is described as a 'pure service', since there is no accompanying good and the production and consumption occur simultaneously. This follows from an understanding of a good as a tangible, alienable product. After all, there seems to be no good that we can point to or buy or sell following a massage.

But while a massage does not result in an alienable product, the service provided by a masseuse does create a physical alteration in the consumer, such as the unbinding of knots and an accompanying relaxation of the back muscles. A new *pattern* emerges in the physical makeup of the consumer as a result of the massage. This change in pattern is valuable; the exchange of money is testament to this. Yet, standard accounts of services fail to incorporate the patterns that many, if not at all, services produce because – while they are *continuants*, in the BFO sense (see section 2.2, below) – they are not alienable; they are not something that is left over after the service is provided that can be bought and sold.

Such patterns are prevalent in every economic sphere; they are the products of many familiar services, which create, maintain, or protect these patterns. A hairdresser might create a new pattern on your head – a product that is neither a good nor a service as standardly understood. A landscape architect might transform a neglected lawn into a well-curated garden. This produces a new pattern in the terrain, though this product, again, cannot be transferred as a good to some third party. A gardener may maintain this pattern by weeding, fertilizing, and trimming the garden. They might lay mulch to protect the garden from future unwanted growth. These acts, too, are typically considered pure services, as no new good has been produced. However, the design of the garden is here not created but rather preserved and protected through the actions of ridding the garden of unwanted growth and introducing measures against future weeds. As it stands, the typical goods-services distinction does not account for this feature of economic production, in which a non-alienable physical change has taken place as a result of the provision of a service.

Yet the coming into existence of products of this sort frequently accompanies the provision of a service, despite not being recognized as an entity that is produced by the service. The principal reason for this, of course, is that the patterns in question are non-alienable. They cannot be bought or sold or gifted, and as a result do not enter, for example, into national income statistics. Yet they are of economic significance, nonetheless. Such patterns include, for example, the patterns in human beings we call skills and capabilities, the sorts of patterns that are created through the provision of those services we call education and training. Taken together the latter form what is called *human capital*, which we discuss further in section 2.3 below.

These patterns are valuable yet underappreciated products of myriad services and, as such, are critical components of market economies. Once patterns are recognized as products, it appears that most, if not all, services derive their value from the value of the resulting patterns that they produce. If this is the case, then patterns in our technical sense – non-alienable products of services – could constitute an important category for assessing and tracking the health of an economy beyond the traditional goods-services continuum.

2.1. Patterns and Health Care Services

In any market economy one important pattern-producing economic sector is health care. In 2021 health care spending in the United States accounted for over 18% of the Gross Domestic Product, and much of the underlying expenditure is directed towards health-related services.[5] For example, in 2017 the Centers for Medicare and Medicaid Services (CMM) estimated that Physician and Clinical Services accounted for about 20% of health care expenditures, or $694.2 billion dollars, where Other Health, Residential, and Personal Care Services accounted for about 5 percent of health care service costs. [6] Many medical services are pattern-producing, typically by altering, maintaining, or protecting patterns that inhere in a patient's body. Curing an infection, analyzing bloodwork, reducing inflammation, and myriad other common medical practices and services result in non-alienable products that promote health and healing. These products – patterns that results from, or are maintained through, medical services, are (arguably) the basic unit of value for assessing such services. For the services are of value only insofar as they are pattern-producing or -protecting in one or other medically relevant sense.

Analyzing such services in terms of patterns provides an alternative framework for assessing and tracking the impact of the health care sector on an economy. Identifying patterns allows for a non-goods-related analysis of the outcomes of health care services – something the goods/services continuum does not provide. And we believe that focusing on patterns offers new insights into, the typical kinds of services that are offered by the medical profession.

Some medical services, such as cosmetic or therapeutic surgery, *create* patterns in patients. If a patient undergoes cleft lip and palate repair surgery, then a new pattern emerges in the patient. The same occurs for breast augmentation surgery or bunion correction. In each case, a service is rendered; no traditional good is produced; yet an important change occurs: a product emerges in the patient. This product is a created pattern. Other services, such as wellness checks from a primary physician, might contribute to the maintenance of patterns. Still others provide protective patterns – patterns which protect against the emergence of unwanted new patterns, as when a physician provides a vaccine against polio. Finally, many services restore patterns to their desired state, such as the removal of a cancerous mole or the treatment of mental trauma through therapy.

The inclusion of references to patterns in the analysis of health care interventions captures an important traceable and analyzable feature of the contributions of the medical profession. The notion of patterns is also useful when services are poorly rendered or abandoned. If services are tracked without reference to patterns, then it becomes possible to track the outcomes of medical interventions in more coherent ways, for example by allowing distinctions to be drawn between services with no ostensible outcome, services with a desired outcome, and services with adverse outcomes.

Consider, for example, coronary bypass surgery. Imagine that three patients are each of them counted as undergoing the service of having this surgery performed. The first has a successful outcome; a new desired pattern emerges in the patient. When the surgeon operates on the second, she finds while operating that a previously undetected issue makes it unfeasible to continue the operation, and so she abandons the surgery mid-operation; new significant knowledge arises, but no new significant pattern results in the patient. In the third patient the surgeon accidentally nicks an artery, causing an adverse outcome; a significant, though unintended, pattern is produced. On its own, the goods/services distinction fails to capture the resulting changes in the patient, where the three different outcomes of the surgeries when understood in terms of patterns can help differentiate between the outcomes and so provide a more systematic framework for assessing value of the respective services.

Many other professions likewise offer services that create, maintain, restore, or protect patterns as products. A car mechanic might manipulate a car's engine, creating a new pattern through their service. They maintain patterns through oil changes, for example, or protect patterns through lubricating the undercarriage of a car. In these cases, the patterns are features of the car. Dentistry is focused on the maintenance, protection, restoration, and (in the case of cosmetic dentistry) creation of teeth. This involves in each case a focus on the pattern that inheres in the jaw or faciocranial features of the patient.

The digital world, too, is host to many services that result in altered patterns. Friedrich's Instagram curator might scrub his account of some photos and modify others – services that result in products not typically categorized as goods but are captured by the idea of patterns. Software developers manipulate code, which as it appears on your hard drive forms patterns in the sense intended here; the job of the software developer is to create, inspect, modify, and protect patterns of code.

Focusing on health outcomes of medical services is, of course, nothing new. Outcome-focused assessments are a widely recognized and powerful tool for evaluating and comparing the contributions of the medical profession within market economies, and service-outcome relations have been explored in terms of ontological concepts[7][8][9]. However, our contribution lies in the recognition that these outcomes are neither goods nor services, traditionally understood, but rather non-alienable products here called patterns.

2.2. Basic Formal Ontology (BFO) and Patterns

Basic Formal Ontology offers the resources to categorize these entities: namely, as dependent continuants. Dependent continuants are entities that exist through time and exist only in virtue of another entity which is their physical bearer – an independent continuant in BFO terms. Examples of dependent continuants include qualities and dispositions, both of which inhere in independent continuants) and are, as such, non-alienable from their bearers [2]. For example, the quality of redness always requires a bearer, such as a tulip or an oil painting; the disposition of fragility requires a bearer, such as a vase. The “dependent” feature reflects the non-alienability of these products, which importantly distinguishes their production from the production of standard goods. We call these dependent continuants patterns. Patterns are the value-adding dependent continuants that are produced by many services; they are immensely prevalent in all economies of production. Thus, in assessing the production of an economy we need to move beyond the standard goods-services distinction in order to reflect the various relations between services and patterns, which may take different forms.

The standard kinds of services include the following:

1) Pure Services: Services that are unaccompanied by any sort of good. For example, a psychiatrist assesses a person’s decision-making capacity, which produces no ostensible good. (She is not required, for instance, to produce any report.) Therapists assist in resolving trauma; occupational and speech therapists change patterns of movement or speech. In all of these cases no good (no alienable continuant) is produced.
2) Goods-Accompanied Services: Services that complement or are required for the use of a good, such as for training or installation purposes. In medicine, examples include the development and use of removable prosthetics or devices such as a Left Ventricular Assist Device (LVAD), which are external, alienable goods that require high levels of service for maintenance and use.
3) Pattern-Producing services: Services that produce enduring ostensible dependent continuants, or patterns, e.g. an Instagram curator produces visible content, a fitness trainer produces enhanced physique, a dentist whitens teeth. In medicine, these might include cosmetic and therapeutic surgeries that result in altered features of the patient’s body.

However, there are also other kinds of services that are valuable on account of their relation to patterns:

4) Pattern-Maintaining services: these are services that *maintain* (though they do not necessarily alter) existing patterns. These might include the services offered by a physician who performs a wellness check, the services provided by a car maintenance mechanic, or a dental hygienist, etc. In these cases, services are provided to ensure that an existing pattern – the healthy human body, in the case of medicine – retains its form.
5) Pattern-Restoring services, or those that *restore* patterns to their desired previous form. This might include landscaping to restore garden beds; a doctor treating a fungal infection; the restoration of an oil painting. Medical services of this sort might include occupational therapies, which restore a patient’s movements to a previous level of functionality.
6) Pattern-Protecting services, or those that protect against the threat of damage or impairment to an existing pattern through the introduction of a distinct service, for example those services offered

by a bodyguard or security firm. An example of such a medical service is the provision of vaccinations against certain diseases or topical creams that prevent rashes from poisonous plants.

As will now be clear, the patterns that are maintained, restored, or protected by services can take many forms and inhere in many different kinds of bearers, whether individual or collective. A pattern can be present in a garden in the form of a design; a pattern can be that of the forces operating within a family in relation to which a family therapist or social worker might intervene; it can be the pattern of forces created when a politician is present in a crowd, the many variations of which may include enhancement through the services of a loudspeaker system or protection by a bodyguard. The resulting patterns are dependent continuants; they are not goods, and so they are not captured by the goods-services distinction, yet they are important value-adding components of services. Identifying which patterns are the products of a given service can explain why that service has value in an economy when seemingly no material goods are at stake.

2.3. Human Capital

An important aspect of any economy is human capital, or the assets of persons that positively contribute to the production process. These assets might include capabilities such as personability, job-related abilities, management skills, networking skills, and so forth. Companies invest directly in developing these assets in employees through education, training, and other programs. The development of human capital, then, is typically understood as a product of pure services, as training modules and educational programs are services with few, if any, related goods.

While economists traditionally recognize human capital as an essential feature of an economy, the nature of human capital and of its relation to goods and services has historically been a matter of debate [10, 11]. Our discussion of patterns and services provides a clear structure for understanding human capital and how services can contribute to its development. These valuable assets of persons are capabilities, where capabilities are (roughly) beneficial dispositions of persons. Like dispositions in general, capabilities are grounded in the physical makeup up of the person who is their bearer. This makeup is a pattern, for example of a neurological sort. When companies invest in creating and maximizing the capabilities of their workers – i.e., their human capital – they are attempting to use services (e.g. training modules) to produce, maintain, or protect certain patterns, which are the grounding of these capabilities. As a person's patterns change through training and education, this person develops new capabilities. Thus, once again: services oriented towards human capital are producing something, namely those patterns which are formed out of capabilities.

2.4. Summary

Nearly all services are directed towards the altering of patterns. Consulting firms review patterns manifested in the behavior of companies, for instance in their manufacturing processes or in their financial distribution processes, and they recommend various ways to alter these patterns to maximize efficiency or minimize costs. A sports psychologist attempts to reorganize an athlete's physiological patterns so that they might respond differently to pressure. A dog trainer alters a dog's neurological patterns so that the dog's responses to the environment change. Sales training is an attempt to reorganize a salesperson's patterns to maximize that person's capabilities to sell goods. Such services do not themselves produce goods; but they do produce results. These results can be described as the production, maintenance, restoration, or protection of patterns. Recognizing and tracking such patterns is particularly important in medicine, as health care services often produce, protect, or maintain patterns in or relating to patients or research subjects, especially in cases where no accompanying good is produced. The full impact of many such health care services is not captured by the typical goods-service distinction; the introduction of patterns-as-products, however, does well to fill this void. The introduction of patterns is thus not merely an academic endeavor; recognizing the various relations between services and patterns can contribute to the assessment of services through the value added by the patterns which they produce.

References

[1] United Nations. Gross Domestic Product Per Capita. https://data.un.org/Data.aspx?d=SNAAMA&f=grID%3A101%3BcurrID%3AUSD%3BpcFlag%3A1

[2] A. Smith, (1776) The Wealth of Nations, with an Introduction by A. Skinner (1969) (London: Penguin Books)

[3] Hill, Peter, ”Tangibles, Intangibles and Services: A New Taxonomy for the Classification of Output.” The Canadian Journal of Economics / Revue Canadienne d’Economique, 32.2 (1999): 426–446.

[4] R. Arp, B. Smith, A. Spear. Building ontologies with Basic Formal Ontology. MIT Press, 2015.

[5] Centers for Medicare and Medicaid Services. NHE Fact Sheet, 2023 URL:https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet

[6] ibid.

[7] Centers for Medicare and Medicaid Services. Quality Measurements and Quality Improvement, 2023 https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/mms/quality-measure-and-quality-improvement-

[8] The Commonwealth Fund: U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes. 2023. https://www.commonwealthfund.org/publications/issue-briefs/2023/jan/us-health-care-global-perspective-2022

[9] M. Fox et al., An Ontological Approach to Analysing Social Service Provisioning, in: 2022 IEEE International Smart Cities Conference, ISC2, Pafos, Cyprus 2022, pp. 1-7, doi: 10.1109/ISC255366.2022.9922132.

[10] J.S. Mill, (1848) Principles of Political Economy, Chapter III. London: Longmans, Green, and Co. 1920.

[11] A. Marshall, (1898) Principles of Economics, Book II. (8th ed.), London: Macmillan and Co. Ltd. 1962

A method to facilitate the identification of FAIRification objectives

César Bernabé [1], Annika Jacobsen [1], Luiz O. Bonino da S.S.[1,2] and Marco Roos[1]

[1] Leiden University Medical Center, Albinusdreef 2, 2333 ZA, Leiden, The Netherlands
[2] University of Twente, Address, Drienerlolaan 5, 7522 NB, Enschede, The Netherlands

1. Introduction

The FAIR principles guide making resources accessible, interoperable, and reusable for humans and machines. The process of making data FAIR (FAIRification) is multi-faceted and can be realised by different means. Consequently, to achieve successful FAIRification, it is essential to define clear objectives that will support establishing a cohesive and effective FAIRification process.

2. Results

We present GO-Plan [1], a method for identifying FAIRification objectives, drawing from experience with recent FAIRification projects (e.g., FAIRification of rare disease data) and feedback from FAIR experts. The method consists of six phases: (i) FAIRification preparation, (ii) assessment of FAIR supporting infrastructure and target resources, (iii) preparation of project collaborators, (iv) identification of domain scope and groups of reuse stakeholders, (v) FAIRification goals refinement and alignment to target FAIR principles, and (vi) decision-making. In the first phase, FAIRification preparation tasks are performed. In phase two, the organisation's infrastructure is assessed for its ability to support initial objectives. In the third phase, stakeholders are categorised into domain and FAIR experts, and knowledge gaps between them are addressed to facilitate communication. In the fourth phase, the domain scope, competency questions for modelling (meta)data, and objectives based on potential *reusers* of the FAIR resource are defined. In the fifth phase, the objectives initially identified are further refined. Finally, implementation decisions are made to achieve the FAIRification objectives in phase six. These phases are accompanied by validation sessions with domain and FAIR experts. The FAIRification objectives identified with GO-Plan are used to guide and constrain the subsequent FAIRification (meta)data modelling steps and to support the selection of ontologies to be reused.

3. Discussion

We anticipate that GO-Plan will help stakeholders better define implementation strategies. By establishing clear objectives through the method, people conducting FAIRification can effectively guide and justify implementation choices and determine the (meta)data elements that must be collected and published to facilitate FAIR data reuse.

4. References

[1] Bernabé et al. A goal-oriented method for FAIRification planning, 23 June 2023, PREPRINT Available at: https://doi.org/10.21203/rs.3.rs-3092538/v1

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: c.h.bernabe@lumc.nl (C. Bernabé); a.jacobsen@lumc.nl (A. Jacobsen); l.o.boninodasilvasantos@utwente.nl (L. O. Bonino da S. S.), m.roos@lumc.nl (M. Roos)
ORCID: 0000-0003-1795-5930 (C. Bernabé); 0000-0003-4818-2360 (A. Jacobsen); 0000-0002-1164-1351 (L. O. Bonino da S. S.), 0000-0002-8691-772X (M. Roos)

CEUR Workshop Proceedings (CEUR-WS.org)

Taxonomy proposal for controlled vocabulary planning in the Pinakes catalogue with Tematres

Josina da Silva Vieira[1], Letícia Santos de Jesus[1] and Tainá Batista de Assis [1]

[1] *Brazilian Institute of Information in Science and Technology (Ibict), Brasília, Brazil*

Abstract

This study presents the results of an experiment to build a taxonomy with the Tematres software to be applied to the Brazilian Integrated Catalog of Bibliographic Records (Pinakes Catalog), is one of the products of the Pinakes Research Project, which aims at restructuring and modernizing the traditional bibliographic services of the Brazilian Institute of Information in Science and Technology (Ibict). The taxonomy encompasses seven categories with their respective subcategories, built from the integration of terms referring to the description of bibliographic records, gathered, organized and disseminated by the information systems of the Bibliodata Network and the National Collective Catalog of Serials (CCN). Through Tematres, the records of this taxonomy can be accessed in various formats, via the web, enhancing their integration and interoperability, highlighting their correlations and maximizing the application of semantic web techniques. The research method includes bibliographic review, document analysis, observation, and survey of terms in the databases of the national information services mentioned. It is expected that the results achieved will contribute to the strengthening of discussions in the area and the opening of new fronts for scientific exploration.

Keywords

Taxonomy, Integration, Interoperability, Tematres Software

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

*Corresponding author.

josinavieiracinf@gmail.com (J. S. Vieira); leticiasantosjesus7@gmail.com (L. S. Jesus); taina@ibict.br(T. B. Assis)

0000-0002-9727-3771 (J. S. Vieira); 0009-0008-7556-9660 (L. S. Jesus); 0000-0002-9176-4764 (T. B. Assis)

CEUR Workshop Proceedings (CEUR-WS.org)

The Interplay of Wikidata and the Cell Ontology

Tiago Lubiana [1]

[1] *University of São Paulo, R. Cidade Universitária, 374 - Butantã, São Paulo, SP, Brazil*

Abstract

This poster presentation centers on the potential of Wikidata, a comprehensive and collaborative knowledge base, to advance biomedical ontologies, particularly the Open Biological and Biomedical Ontologies (OBO) Foundry ontologies. The spotlight is on the integration of Wikidata with the ontology landscape, showcasing its integration with the Cell Ontology (CL).The CL, established around 2005, has been expanding to currently include over 2,600 cell classes, with technical enhancements and the addition of novel cell types over the years. Despite these advancements, contributing to it can be daunting, and its coverage of cell types is yet far from complete.Ontologies, though highly effective, require substantial technical expertise and specific software installations, restricting potential contributors with limited access or skills.In contrast, Wikidata promotes communal contribution and is welcoming for users of varying technical proficiency. Its strength stems from a vast community of active contributors, with broad coverage of diverse concepts. The speed, breadth, and accessibility of Wikidata, along with its CC0 public domainlicense, render it an inviting platform for widespread use and contribution.Our work exemplifies the advantages arising from the integration of Wikidata with OBO Foundry ontologies. For instance, Wikidata's multilingual capability promotes inclusivity in stark contrast to the predominantly anglocentric ontologies and, its direct linkages to Wikipedia facilitate access to textual descriptions often lacking in ontology development. Finally, Wikidata's connections to Wikipedia and other controlled vocabularies can support ontologies in validating information.To map concepts in Wikidata to the Cell Ontology, we developed a streamlined workflow combining ROBOT and a custom Python package for real-time natural language term curation in Wikidata, resulting in over 2,600 cross-references from Wikidata to the Cell Ontology. These curation efforts have revealed several cell types listed on Wikipedia but absent from CL, providing a unique opportunity for expansion and enhancement of CL's scope, as well as opening a gateway for multilingualism in applications that use CL identifiers.To sum up, the integration of Wikidata and ontologies can significantly benefit the ontology ecosystem, enriching the value of ontologies and fostering broader collaboration. Our work with the Cell Ontology suggests its connection other ontologies is a promising work direction for the future.

Keywords

Wikidata, Cell Ontology, cell types, knowledge graph.

1. Acknowledgements

This work was funded by FAPESP grant #19/26284-1.

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: tiago.lubiana.alves@usp.br (A. 1)
ORCID: : 0000-0003-2473-2313 (A. 1)

Tracking the functional effects of SARS-CoV-2 genomic variants: An ontology-driven approach

Madeline Iseminger[1,2,*], Muhammad Zohaib Anwar[2], Rhiannon Cameron[2], Damion Dooley[2], Paul Gordon[3], Emma Griffiths[2], Anoosha Sehar[2], Khushi Vora[3] and William Hsiao[1,2]

[1] *University of British Columbia, Vancouver, BC, Canada*

[2] *Centre for Infectious Disease Genomics and One Health, Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada*

[3] *Centre for Health Genomics and Informatics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada*

Abstract

Emerging SARS-CoV-2 genomic variants can impact disease transmission, viral antigenicity, infection severity, and vaccine efficacy. As such, it is critical that new variants and their potential impacts are tracked in a rapid and globally accessible way. Due to the intensive labor required to manually extract genomic variant information from the literature, a semi-automated approach is needed. We present a novel ontological framework for describing SARS-CoV-2 mutations and their purported functional effects, and contextual data for literature evidence. This framework follows Basic Formal Ontology guidelines and is interoperable with existing OBOFoundry ontologies. When coupled with genomic surveillance of circulating pathogens, it will assist with rapid sharing of potential functional impacts of new variants in a standardized, machine-readable way. In future, the model could be extended to use cases beyond SARS-CoV-2, such as influenza or antimicrobial resistance. The framework consists of three linked minimodels: variant calling, host and pathogen phenotypes, and literature evidence. The variant calling model describes the process from sequencing a viral sample to variant calling, and linking variant calls to phenotypes. As far as we know, this is the first model in OBOFoundry to describe mutation-phenotype relations. The mutation names exist on the instance level to avoid proliferation of new classes, and they are correlated with punned instances of phenotypes. Sequence Ontology[1] terms for mutation types were not used to remain compatible with BFO standards. The phenotype model contains terms for functional impacts that are correlated with SARS-CoV-2 mutations, spanning levels of granularity from molecular impacts to impacts on disease transmission. The terms are based on terms from Pokay[2], a hand-curated repository of SARS-CoV-2 mutations and their functional effects, with links to related research articles. The phenotype terms are housed in the Pathogen Host Interaction Phenotype Ontology (PHIPO)[3], while non-phenotype terms (relating to vaccines, treatment, diagnostics, and associations with pre-existing conditions or homoplasy) are reused from the Vaccine Ontology (VO)[4], Coronavirus Infectious Disease Ontology (CIDO)[5] wherever possible. Phenotype terms begin with "altered", matching PHIPO[3], as a description of the change taking place. The literature evidence mini model links mutation calls and phenotypes to their literature evidence sources. Short free text descriptions of the research findings are included here. We are in the process of developing a text mining module utilizing the minimodels to explore semi-automatic retrieval of relevant literature.

Keywords

SARS-CoV-2, application ontology, mutations and phenotypes, literature retrieval

References

[1] K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The sequence ontology: a tool for the unification of genome annotations, Genome Biol. 6 (2005) R44.

[2] P. Gordon, Pokay, https://github.com/nodrogluap/pokay, ???? Accessed: 2023-6-7.

[3] M. Urban, A. Cuzick, J. Seager, V. Wood, K. Rutherford, S. Y. Venkatesh, J. Sahu, S. V. Iyer, L. Khamari, N. De Silva, M. C. Martinez, H. Pedro, A. D. Yates, K. E. Hammond-Kosack, PHI-base in 2022: a multi-species phenotype database for Pathogen-Host interactions, Nucleic Acids Res. 50 (2022) D837–D847.

[4] Y. Lin, Y. He, Ontology representation and analysis of vaccine formulation and administration and their effects on vaccine immune responses, J. Biomed. Semantics 3 (2012) 17.

[5] Y. He, H. Yu, E. Ong, Y. Wang, Y. Liu, A. Huffman, H.-H. Huang, J. Beverley, J. Hur, X. Yang, L. Chen, G. S. Omenn, B. Athey, B. Smith, CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis, Sci Data 7 (2020) 181.

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

*Corresponding author.

m.iseminger@alumni.ubc.ca (M. Iseminger)

0000-0002-0548-891X (M. Iseminger); 0000-0001-8236-485X (M. Z. Anwar); 0000-0002-9578-0788 (R. Cameron); 0000-0002-8844-9165 (D. Dooley); 0000-0002-1107-9135 (E. Griffiths); 0000-0001-5275-8866 (A. Sehar); 0000-0002-1342-4043 (W. Hsiao)

Ontology Development Strategies and the Infectious Disease Ontology Ecosystem

Giacomo De Colle [1], Ali Hasanzadeh[1] and John C. Beverley [1,2]

[1]*University at Buffalo, Buffalo, NY, USA*
[2]*National Center for Ontological Research, Buffalo, NY, USA*

Abstract

After motivating a framework for evaluating top-down, middle-out, middle-in, and bottom-up ontology development strategies, we apply our framework to investigate whether infectious disease ontologies - specifically, the Virus Infectious Disease Ontology (VIDO) and the Coronavirus Infectious Disease Ontology (CIDO) - effectively promote semantic interoperability.

Keywords

Top-down, middle-out, bottom-up, middle-in, infectious disease ontologies, CIDO, VIDO

1. Introduction

Ontologies are developed using numerous strategies. Some follow a *top-down* strategy, in which classes or categories are devised to constrain lower-level ontology content extending from them [1]. *Bottom-up* strategies tend to begin creating ontology content reflecting a given domain of interest, representing that content with a high degree of fidelity [2]. Data used as a basis for a bottom-up ontology comes in many flavors, e.g. a SQL database, an Excel file, previously existing taxonomies, etc. An ontologist following the bottom-up strategy will attempt to uncover terms and relations implicit in the data, to represent them ontologically. The *middle-out* strategy aims at attaining the benefits of the preceding strategies, like proximity to the domain and ensuring consistency across lower-level ontologies [3,4]. Middle-out ontologies are developed at some level of abstraction above one or more domains to be modeled, but not as far removed as those that begin with the *top-down* strategy. Notice that these strategies are distinguished based on the starting point of their development [5]. They are thus distinct from ontology architectures distinguished in terms of their coverage, e.g. top-level ontologies such as the Basic Formal Ontology (BFO) [1,6], DOLCE [7], and YAMATO [8]; *mid-level* ontologies – such as the Common Core Ontologies (CCO) suite [9] or the Industrial Ontology Core [10].

While each of these strategies has been discussed in the literature, there has, as of yet, not been a rigorous, fair, comparison provided between them. One of our aims in this article is to provide such a comparison. Another of our aims is to apply the results of our comparison to representative ontologies which follow one of these strategies, specifically, the Coronavirus Infectious Disease Ontology (CIDO) [11,12] and the Virus Infectious Disease Ontology (VIDO) [13], designed according to the top-down strategy. Our evaluation will demonstrate the extent to which these ontologies effectively promote semantic interoperability, a primary goal of ontology development.

2. Criteria for Evaluating Ontologies

[1]Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: gdecolle@buffalo.edu (A. 1); ahasanza@buffalo.edu (A. 2); johnbeve@buffalo.edu (A. 3)
ORCID: 0000-0002-3600-6506 (A. 1); 0009-0003-6686-3319 (A. 2); 0000-0002-1118-1738 (A. 3)

We take as a starting point for the identification of evaluative criteria for ontologies, the work of Denny Vrandečić [14], itself originated from the work of Thomas Gruber. Vrandečić's evaluating criteria – displayed in Table 1 – have been used by numerous developers to improve the quality of their ontologies and knowledge graphs [15-7]. Note, no single ontology can perform well on all these metrics; indeed, some criteria appear in conflict, such as *completeness* and *conciseness*.

Table 1
Vrandečić's Ontology Evaluation Criteria

Criterion	Description
Accuracy	The extent to which an ontology accurately represents the domain within its stated scope.
Adaptability	The extent to which an ontology can be extended to represent entities in domains outside of its originally stated scope.
Clarity	The extent to which an ontology unambiguously, clearly, conveys the meanings of its terms and relations to users.
Completeness	The extent to which an ontology includes terms and relations that cover the entire domain within the intended scope of the ontology.
Conciseness	The extent to which an ontology is parsimonious, does not include redundant content, or irrelevant axioms.
Coherence	The extent to which an ontology is both logically consistent and semantically aligned with the intention of its creators.
Organizational fitness	The extent to which, within an organizational context, an ontology is integrated within the organization.

Accuracy follows naturally from the goals of most ontology development, namely, representing a given domain using a machine-readable controlled language. Accuracy may be determined when ontology developers and subject-matter experts interact during ontology development, in the interest of reaching consensus over ontology labels, definitions, logical relationships, etc. [18]. The most successful ontologies are reused, extended to new domains, and integrated with other ontologies and knowledge representation projects. An *adaptable* ontology is one which can be easily extended into a new domain distinct from that for which it was initially designed. *Clarity* can be achieved by proper definition development and documentation. In the absence of definitions, other annotation properties such as comments, citations, notes, labels, alternative labels, and preferred labels may be used to promote clarity. A *complete* ontology will include terms and relations needed to represent any terms within its scope. Completeness is related to accuracy, as evidenced by the impact completeness may have on whether competency questions for an ontology can be answered [19]. Adequately answering competency questions requires that the ontology has adequate coverage of the domain. A *concise* ontology will not include unnecessary elements or axioms, which promotes understanding by users and helps avoid the confusion that might emerge from the presence of, say, many unneeded classes. Related, a concise ontology will include only those terms and relations that are needed to represent the domain within its scope, i.e., the minimal set of terms and relations. A *coherent* ontology will be logically consistent and will entail as little beyond the intent of its creators as possible. Put another way, the intended interpretation of the ontology will match as closely as possible the semantic interpretation generated by model checkers [20] or OWL reasoners [21,22]. *Organizational Fitness* is not directly about a given ontology *per se* but also involves the organizational context in which the ontology is deployed. An ontology that scores high on organizational fitness is successfully and consistently deployed in an organizational context, is being maintained and developed, is accessible to members of the organization, and can be aligned with other ontologies within the organization.

The standard we use to adjudicate the importance of each of these criteria is, arguably, the primary aim towards which any good ontology aims, the promotion of *semantic interoperability* – the ability of computer systems to exchange data with unambiguous, shared meaning [1,23-4]. The most important criteria in this perspective seem to be accuracy, coherence, and adaptability. Notice that *coherence* cannot be simply reduced to the absence of contradictions in the ontology but is rather the absence of statements that are semantically, and often implicitly, in conflict. During the development of VIDO, such a coherence issue was identified with respect to the definition of "organism" used by the Ontology for Biomedical Investigations (OBI), which included viruses within purview [25]. "Organism", however, was defined as comprised of cellular entities, despite viruses being acellular. This is an example of a semantically incoherent statement that was originally not noticed and that might be more easily avoided by adopting one ontology development strategy over the others.

3. Evaluating Ontology Development Strategies

The three strategies in our focus are evaluated on a spectrum. The evaluation of these strategies was carried out by analyzing ontologies that explicitly adopt them, and then testing them against a set of questions devised to check compliance with the our criteria. In the future, we plan to test these questions and criteria on a larger scale using empirical methods, i.e. surveys and feedback group sessions with ontology developers, etc. In what follows, we do not conclude that ontologies borne from the one strategy necessarily perform better than others with respect to our criteria. Instead, we suggest that they stand a better chance of doing so than ontologies developed following one of the other strategies.

Organizational Fitness

- **Bottom-up:** This strategy promotes work on the same data set. However, different segments of this dataset, when used by various team members, might lack semantic uniformity. This disparity becomes apparent when generalizing the embedded knowledge.
- **Middle-out:** This strategy provides a common starting point for building classes that members of an ontology development team can use. Nevertheless, there may be a lack of shared understanding across ontologies borne out of this strategy within an organization, in particular with respect to higher-level terms and relations such as **part of**, **quality**, or **process**.
- **Top-down:** While starting from the most general classes and working your way down is a suitable way to promote consistency, it is not free of challenges. Deciding how best to define such classes across an enterprise requires time and skill that organizations may not have.

Definitions (Accuracy, Clarity, Coherence)

The Aristotelian scheme for definition writing is standardly employed by ontology developers [1,26], i.e., to define class A you identify its parent class B and describe the differentia that distinguishes instances of A from any other instances of B [27]. Adopting the Aristotelian model forces the ontologist to identify the relation of the class with the other classes in the hierarchy when building the definition, and provides a format that is human-readable, consistent, and understandable for the user if correctly applied.

- **Bottom-Up:** One cannot adopt the model of writing Aristotelian definitions if following this strategy, since there is no top-level ontology from which to identify parent classes. Consequently, those following the bottom-up strategy must develop some manner of schema to do so. Whatever that schema amounts to, it will not be extending from a top-level ontology, which seems a cost.
- **Middle-Out:** The middle-out strategy may adopt the Aristotelian schema for definitions to some extent, i.e., for classes on the lower level. Nevertheless, it is not possible to adopt this strategy for classes on the upper level of the hierarchy.
- **Top-Down:** Insofar as the top-down strategy allows for extensive application of Aristotelean definition schema, it scores highly with respect to definitions. By using the Aristotelian method, ontology developers can offer definitions by relating them to classes in a top-level ontology.

Axiomatization (Accuracy, Clarity, Consistency)

Formal axioms associated with terms and relations in an ontology promote accuracy, clarity, and consistency, as well as automated checking of such evaluative criteria. Axioms provide machine-interpretable enforcement of domain and range for relations, ensure disjoint classes do share instances in common, and in general connect parts of an ontology hierarchy to other parts of that hierarchy.

- **Bottom-Up:** Axiom development may be challenging when following the bottom-up strategy. Suppose a class **start time** should be related to a class **earlier than** that itself relates to any class representing entities having a duration, such as **phosphorylation** or **dimerization**. Intuitively, one might say such classes are all activities or processes of some sort. This is not an obvious option when following the bottom-up strategy.
- **Middle-Out:** The middle-out strategy fares better, since axioms can be written starting from a broader scoped architecture than that provided by the bottom-up strategy. This allows those following a middle-out strategy to leverage the space of possibilities already constrained by the placement of lower-level terms within the ontology taxonomy. Put another way, following the middle-out strategy might avoid some of the issues raised above by including natively a class-like **activity**, but will at some point lack a way to answer questions further up the taxonomy.
- **Top-Down:** The top-down strategy appears to fare better than either of the preceding strategies, with respect to axiom development. By populating ontology entities downward from existing classes and relations, ontology elements have an implicit formal structure inherited from the top-down strategy itself. Moreover, because there are an increasing number of logical constraints enforced as one proceeds down the taxonomy, the scope of possible axioms that can be applied to an ontology element is decreased.

Reinvention (Accuracy, Organizational Fitness, Clarity, Consistency)

It is important to determine the extent to which an ontology strategy encourages or discourages duplication of effort. Duplication is not only dangerous because it wastes time and resources. It also risks re-creating the mistakes that other ontologists have effectively amended in their efforts.

- **Bottom-Up:** The bottom-up strategy appears to encourage duplicative effort when viewed from the perspective of interoperability. Focusing solely on accurately modeling a domain runs the risk of missing the forest for the trees. Put another way, creating terms and relations highly specific to a domain, without reflection on how they might relate to existing ontologies, impedes interoperability with those other ontologies. This is, indeed, a recipe for creating data silos [1].
- **Middle-Out:** Middle-out ontologies avoid some of the issues plaguing the bottom-up strategy with respect to reinvention, by facilitating an upwards population of ontology content, as needed. Nevertheless, this strategy will ultimately run into the same issues at a higher level of generality.
- **Top-Down:** Top-down strategies clearly shine with respect to reinvention. This strategy creates a common architecture from which terms and relations extend, and consequently, can be reused in other ontologies employing the same top-level architecture. This is, of course, not to say that all ontologies designed according to this strategy thereby perform well with respect to reinvention. Several ontologies, for example, in the Open Biological and Biomedical Ontology (OBO) Foundry [28] extend from the top-level BFO but in the absence of collaboration across nearby efforts, do so by producing duplicative content.

Unused Content (Conciseness, Clarity)

In some cases, ontologies will be developed with placeholder classes or relations, intended to be connected to data, but which never are. The result is there may be ontology classes that are not used, potential points of confusion, or perplexity for ontology developers unfamiliar with the intent behind creating such content.

- **Bottom-Up:** Bottom-up strategies perform best with respect to this criterion, as ontology content is developed on demand, directly from relevant domain data. Bottom-up ontology development encourages creating ontology content that is representative of the domain, and so makes it unlikely that ontology developers following this strategy will create empty content.
- **Middle-Out:** Middle-out strategy followers arguably perform as well as bottom-up strategy followers with respect to this criterion. By keeping an eye on the domain-level and the upper-level, this strategy tends to result in ontologies that populate classes which are integrated into analyses of the domain in question.
- **Top-Down:** Top-level strategy results perform worst with respect to this criterion. Constructing ontologies from the most general content down runs the risk of including plausible content that does not end up being used by domain ontologies extending from the top-level.

Abstract Representations (Clarity, Conciseness, Accuracy)

Ontologies sometimes include content that is so general that may lead to confusion by subject-matter experts, ontology developers, or other users. Such abstract representations may be challenging for users to understand, which may, in turn, undermine the use of the ontology, or support perceptions that a given ontology is too difficult to be used for practical purposes.

- **Bottom-Up:** More than any of the alternative strategies, bottom-up strategies typically avoid the inclusion of abstract representations within ontology output. Working so closely at the domain level encourages the creation of ontology content understood by relevant subject-matter experts.
- **Middle-Out:** Similarly, the middle-out strategy promises to avoid the inclusion of many abstract representations, given its emphasis on domain-level content while attending to top-level architecture. Expanding upwards only when needed keeps abstract representations at a minimum.
- **Top-Down:** Abstract representations are often found in ontologies designed following a top-down strategy, as top-level content must often be rather general. Terms such as **continuant** or **predicate** are often divorced from the experiences of subject-matter experts, employment of them leads to confusion at best and misuse at worst. Consequently, it becomes challenging to link these abstract representations to domain-level content, which is indeed one of the main purposes of developing ontologies.

Adaptability (Adaptability, Organizational Fitness)

Ontologies should be extendable to new domains, reflecting new discoveries, scientific advancements, or novel ways of understanding existing knowledge. We must take care here, however, when reflecting on what it means for an ontology development strategy to promote adaptability. Ontologies may be extended vertically - by upward or downward population of content – or horizontally – by covering new domains not yet represented. The upward and downward population of content places constraints on the ontology content that can be created consistently. For example, asserting that instances of **process** must have some **temporal part** requires that any subclasses of **process** also have some **temporal part**, thereby narrowing the space of possible extensions. A horizontal extension is, however, sometimes entirely unconstrained, as when an ontology is developed outside of existing ontologies.

- **Bottom-Up:** Strictly speaking, this strategy promotes adaptability solely in the sense of horizontal extension since it encourages the creation of ontology silos, entirely disconnected from other potential ontologies within an enterprise. New terms and relations can be introduced outside the scope of existing ontologies developed following this strategy, without encountering inconsistency. This, of course, comes at the cost of interoperability.
- **Middle-Out:** This strategy fares poorly when adaptability is understood as a vertical extension since constraints are applied during the upward and downward population of ontology content. Ontology content developed in the upward direction must also bear constraints from the lower-level ontology content, and vice versa. The upward and downward aspects of this strategy lead it thus compromised on two fronts. Consider, from the Credential Transparency Description Language [29]: an **address** is defined as "particulars describing the location of the place", whereas 'particular' refers to details of a description. Such use of 'particular' is, however, wildly different from the use of 'particular' in, say, DOLCE [5], where the term denotes an individual. Extending Credential Engine to a domain covered by a DOLCE-based ontology would thereby generate inconsistency. On the other hand, where adaptability is understood as horizontal extension the middle-out strategy permits the development of ontologies with few constraints, with the caveat simply being that at some point such new ontologies will need to connect with a top-level. In this respect, the middle-out strategy provides flexibility in ontology design.
- **Top-Down:** The top-down strategy encourages constraints on ontology content extended from the top-level, and so constrains the range of possible ontology extensions. Put another way, when adaptability is understood as a vertical extension, the top-down strategy places constraints on a downward population of ontology content and does not permit an upward population. In this respect, the top-down strategy fares worse than the bottom-up strategy with respect to permitting extensions without generating inconsistency, since any downward populated content

must remain consistent with the content from which it extends. The top-down strategy fares worse, moreover, than either of the other strategies when considering adaptability from the horizontal perspective. This is because any ontology designed following the top-down strategy must find parent terms and relations that ultimately have roots in a top-level ontology. In other words, when attempting to cover a new domain, ontology developers must *find* constraints to apply to ontology content for this new domain.

Implicitness (Accuracy, Consistency, Coherence)

It is important to be able to discern the extent to which an ontology design strategy encourages or discourages the creation of implicit incoherence.

- **Bottom-Up**: While the bottom-up strategy offers developers greater flexibility, this approach has its challenges when it comes to inviting implicit incoherence. The significant dependence on data often results in classes that mirror imperfections in the data. Consequently, terms and relations falling out of a bottom-up strategy tend to reflect inherent flaws in the data.
- **Middle-Out**: Without an explicit structure guiding expansion, ontology developers following this strategy may overlook the "implicit rules" essential for coherence. Additionally, the absence of a shared foundational understanding can render much of the stored knowledge inconsistent, without obvious methods for detecting such inconsistency. Moreover, transitioning the ontology to different domains can be challenging for middle-out strategy ontologies, especially when the abstract classes it has developed are too domain-specific. As discussed, the term **address** in the Credential Transparency Description Language [29] is defined differently than in other ontologies, which can lead to confusion when seeking alignment.
- **Top-Down**: This strategy exhibits advantages and challenges. The top-down strategy equips the ontologist with the tools needed to craft robust and rigorous axioms, paired with clear class definitions, reducing the risk of implicit inconsistencies. However, the rigidity of the top-level classes can be a limitation. If incoherence arises, rectifying it can be challenging, especially if it necessitates changes to foundational classes.

4. Discussion

This section offers a comprehensive analysis of three ontological strategies, summarizing their strengths, weaknesses, and unique characteristics with respect to the preceding evaluative criteria. We aim to provide a holistic view, facilitating a deeper understanding of each method's applicability and limitations.

The bottom-up strategy excels in clarity and conciseness, primarily because the ontologies it produces are more closely aligned with the terminology associated with a given domain. When conducted effectively, this strategy accurately mirrors domain data structure, ensuring a direct correlation with the original data's knowledge. However, such close reliance on original datasets can compromise adaptability and accuracy when applied to slightly varied domains. The bottom-up strategy’s biggest strength - its proximity to the original data – is also its biggest weakness.

The middle-out strategy generally outperforms the bottom-up method. Ontology developers following this strategy tend to avoid getting mired in data details. Nevertheless, the absence of a top-level set of common terms can lead to implicit commitments in an ontology developed following the middle-out strategy. Moreover, developers will ultimately be pressed in this strategy to create local, domain-dependent top-level terminology, that will quite likely impede clarity and adaptability, and lead to challenges in semantic homogeneity and external interoperability.

The top-down strategy stands out in terms of coherence, clarity, and completeness. It offers adaptability, especially when using top-level classes to ensure coherence across domains. However, its inherent ontological commitments may make it fragile, necessitating careful adaptation to new domains, and rigorous attention to how ontologies are extended from it. While the approach promotes ontology alignment and avoids redundancy, it demands a centralized organizational structure and skilled individuals adept at handling abstract classes.

The top-down strategy appears to perform best overall with respect to our evaluative criteria, with the middle-out strategy coming second, and the bottom-up strategy third. While challenges exist, the top-down method's proactive approach to axiomatization and class building during production can also mitigate issues that might arise post-deployment. That said, it is important to emphasize that applying top-down techniques often requires more effort in the development of an ontology. Compliance to a top-level is not cheap, and the benefits gained are not always obvious when an ontology is being developed for a specific domain-level modeling task. The tradeoff is – we claim – best characterized as a difference between short-term and long-term costs. Effort spent upfront following a top-down strategy saves effort downstream, a cost that artifacts resulting from middle-out and bottom-up strategies will ultimately have to pay later – when it is more expensive - if they hope to promote semantic interoperability.

5. The Middle-In Strategy

Most researchers appear to develop ontologies according to one of the three developmental strategies discussed thus far, but our analysis of the complexity of ontology development suggests there is a promising strategy not yet discussed in the literature. The *middle-in* strategy combines aspects of the top-down and bottom-up strategies, in a manner distinct from the middle-out strategy. According to this strategy, ontologists begin with a top-level ontology which is then used as a guide when exploring data to develop domain-level ontology content. This strategy takes its name from starting at both the top and bottom levels, then developing ontology content to meet in the middle.

During the writing of this manuscript, it was discovered that some ontology developers have – often unknowingly – employed such a strategy or have hinted towards such a strategy. For example, in 2010 Enrico Francesconi and his team published a paper where they describe a similar method to build DALOS, a multilingual ontology for the legal domain [30]. More recently, CIDO and VIDO are the result of the middle-in strategy, as they each use BFO as a top-level ontology, while employing bottom-up, data-focused, design. For example, they re-use reference ontologies and mid-level ontologies like CHEBI [31] and OBI [25]. Moreover, CIDO was built starting from the classifications of real-life, already existing data coming from GISAID, NextStrain, and DrugBank, as well as data coming from domain-specific literature [32], and VIDO was related to taxonomies such as the NCBITaxon [33] and based on the Baltimore Classification [34]. Finally, both made use of connections with domain experts to maintain their terminology and ontological commitments grounded within the domain. The perks of employing top-down methods are still visible in the quality of employed definitions and axioms, as well as in the number of external ontologies and domains that are referred to, suggesting the ontologies are adaptable and extensible.

CIDO and VIDO score well on all criteria most relevant for semantic interoperability. Starting with accuracy, the precision of the content of the two stems from the quality and variety of the data and domain knowledge used as a development basis. Clarity is favored by the terminology respecting these sources, as well by the terms being defined using the Aristotelian schema. Coherence is respected through the use of axioms taken or developed starting from the top-level layer provided by BFO. Adaptability is preserved thanks to native integration with a set of BFO-based ontologies and to potential coherent integration with all other ontologies that use the same upper-level architecture. These virtues make CIDO and VIDO stable hubs for long-term development of interoperable terminologies in the realm of infectious disease representation, and provide confirmation for our evaluation of middle-in ontologies as best suited to promote semantic interoperability.

6. Conclusion

Each of the four methodologies presents distinct advantages and challenges. The bottom-up approach, with its implicit semantics, is apt for smaller projects where team members share a common understanding and where rapid access to existing well-defined datasets is needed. Conversely, for larger projects demanding explicit semantics to maintain coherence across multiple contributors, extract implicit information, or promote reasoning capabilities, the middle-out and top-down approaches are more appropriate. Both, however, are less appropriate than the middle-in strategy which concluded our discussion. In the context of infectious disease ontologies, CIDO and VIDO represent successful endeavors in providing a basis for structuring data. We argue that the quality of the two ontologies is in part a function of their adoption of a middle-in development strategy. While the middle-in strategy is akin to the top-down strategy, the former may yield more fragile ontologies due to its rigorous nature. Proper implementation demands meticulous effort from ontologists in crafting precise mappings and alignments. As we understand it the primary impediment to interoperability arises from inconsistent terminology and axioms across domains; this is an issue best addressed by the middle-in strategy.

Acknowledgements

The authors wish to acknowledge the help and comments of Alexander Diehl, Asiyah Lin, the members of the Spring 2023 UB Logic for Ontology seminar, the PROVO-BFO mapping group, the 2023 ICBO Infectious Disease workshop audience and especially of one of its participants, who pointed out that empirical tests should accompany our current methodology for testing adherence to the criteria.

References

[1] Arp Robert, Spear Andrew D., Smith Barry, “Building Ontologies with Basic Formal Ontology”, MIT Press, 2015.

[2] Keet, Maria C. “Introduction to Ontology Engineering”, College Publications, 2020.

[3] Belhoucine, K., Mourchid, M., Mouloudi, A., Mbarki, S., "A Middle-out Approach for Building a Legal domain ontology in Arabic," 2020 6th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, (2020): 290-295.

[4] Uschold Mike, Grüninger Michael. “Ontologies: Principles, Methods and Applications.”, The Knowledge Engineering Review, vol. 11, no. 2, pp. 93–136 (1996)

[5] Raad Joe, Cruz Christophe, “A survey on ontology evaluation methods”, Proceedings of the International Conference on Knowledge Engineering and Ontology Development, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, (2018).

[6] BFO 2020, 2020, URL: https://github.com/BFO-ontology/BFO-2020

[7] Borgo Stefano, Claudio Masolo. “Ontological Foundations of Dolce”, Theory and Applications of Ontology: Computer Applications, pp. 279–295, (2022).

[8] Mizoguchi Riichiro, Borgo Stefano, “Yamato: Yet-Another More Advanced Top-Level Ontology”, Applied Ontology, vol. 17, no. 1, pp. 211–232, (2022).

[9] CCO, Common Core Ontologies 2023, URL : https://github.com/CommonCoreOntology/

[10] Kulvatunyou, B. , Drobnjakovic, M. , Ameri, F. , Will, C. and Smith, B. (2022), “The Industrial Ontologies Foundry (IOF) Core Ontology”, Formal Ontologies Meet Industry (FOMI) 2022, Tarbes, FR, (2022).

[11] Yongqun He, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, anthony Huffman, Hsin-hui Huang, John Beverley, Junguk Hur, Xiaolin Yang, Luonan Chen, Gilbert S. Omenn, Brian Athey and Smith Barry, “CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis”, Scientific Data Nature, (2020).

[12] He Yongqun, Yu Hong, Hufman Anthony, Yu Lin Asiyah, Natale Darren A., Beverley John, Zheng Ling, Perl Yehoshua, Wang Zhigang, Liu Yingtong, Ong Edison, Wang Yang, Huang Philip, Tran Long, Du Jinyang, Shah Zalan, Shah Easheta, Desai Roshan, Huang Hsin-hui, Tian Yujia, Merrell Eric, Duncan William D., Arabandi Sivaram, Schriml Lynn M., Zheng Jie, Masci Anna Maria, Wang Liwei, Liu Hongfang, Smaili Fatima Zohra, Hoehndorf Robert, Pendlington Zoë May, Roncaglia Paola, Ye Xianwei, Xie Jiangan, Tang Yi-Wei, Yang Xiaolin, Peng Suyuan, Zhang Luxia, Chen Luonan, Hur Junguk, Omenn Gilbert S., Athey Brian and Smith Barry, “A comprehensive update on CIDO: the community-based coronavirus infectious disease ontology”, Journal of Biomedical Semantics, (2022).

[13] Beverley, John ; Babcock, Shane ; Smith, Barry ; He, Yongqun ; Merrell, Eric ; Cowell, Lindsay ; Hurley, Regina & Duesing, Sebastian (2022). Coordinating Coronavirus Research: The COVID-19 Infectious Disease Ontology. Proceedings of the International Conference on Biomedical Ontologies.

[14] Vrandečić Denny, “Ontology Evaluation.” in: Steffen Staab, Rudi Studer (Ed.), Handbook on Ontologies, 200, pp. 293–313.

[15] Diehl Alexander D., Lee Jamie A., Scheuermann Richard H., Blake Judith A., “Ontology development for biological systems: immunology”, Bioinformatics Applications Note, Vol.23 no.7 (2007): 913-915.

[16] Hogan, Aidan, et al. "An empirical survey of linked data conformance.", Journal of Web Semantics 14, pp. 14-44, (2012).

[17] Zaveri Amrapali, Rula Anisa, Maurino Andrea, Pietrobon Ricardo, Lehmann Jens, Auer Sören, "Quality assessment for linked data: A survey", Semantic Web 7.1 pp. 63-93, (2016).

[18] Neuhaus Fabian, Hastings Janna, "Ontology is Consensus Creation, Not (Merely) Representation, Applied Ontology 17, (2022).

[19] Fox Mark, Grüninger Michael, "The Role of Competency Questions in Enterprise Engineering", Proceedings of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, (1994).

[20] McCune William, "Mace4 Reference Manual and Guide", Argonne National Laboratory Technical Memorandum n.264, (2003).

[21] Hermit, an OWL 2 reasoner, 2023. URL: http://www.hermit-reasoner.com/

[22] Pellet, an OWL 2 reasoner, 2023. URL: https://github.com/stardog-union/pellet

[23] Abukwaik H, Taibi D, Rombach D (2014) Interoperability-related architectural problems and solutions in information systems: a scoping study. In: Avgeriou P, Zdun U (eds) Software architecture. ECSA 2014. Lecture notes in computer science, vol 8627. Springer, Cham. https://doi.org/10.1007/978-3-319-09970-5_27

[24] Justine Flore Tchouanguem, Mohamed Hedi Karray, Bernard Kamsu Foguem, Camille Magniont, F. Henry Abanda, and Barry Smith. 2021. BFO-based ontology enhancement to promote interoperability in BIM. Appl. Ontol. 16, 4 (2021), 453–479. https://doi.org/10.3233/AO-210254

[25] Bandrowski A, et al. (2016) *The Ontology for Biomedical Investigations*. PLoS ONE 11 (4): e0154556. doi: 10.1371/journal.pone.0154556.

[26] Otte NJ, Beverley J, Ruttenberg A. (2022) *BFO: Basic Formal Ontology*. Applied Ontology. 17-43.

[27] Seppälä Selja, Ruttenberg Alan, Schreiber Yonatan, Smith Barry, "Definitions in ontologies", Cahiers de Lexicologie 109 (2), pp.175-207, (2016).

[28] Smith B, et al. (2007) *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol. 25:1251–1255. doi:10.1038/nbt1346.

[29] Credential Transparency Description Language (CTDL), 2023, Accessed 8-14-23 from https://credentialengine.org/credential-transparency/ctdl/

[30] Francesconi Enrico, Montemagni Simonetta, Peters Wim, Tiscornia Daniela, "Integrating a Bottom–UP and Top–down Methodology for Building Semantic Resources for the Multilingual Legal Domain", Semantic Processing of Legal Texts, pp. 95–121, (2010).

[31] Hastings J. et al. *ChEBI in 2016: Improved services and an expanding collection of metabolites*. Nucleic acids research 44, D1214-1219, doi:10.1093/nar/gkv1031 (2016).

[32] Yu Hong, Li Li, Huffman Anthony, Beverley John, Hur Junguk, Merrell Eric, Huang Hsin-hui, Wang Yang, Liu Yingtong, Ong Edison, Cheng Liang, Zeng Tao, Zhang Jingsong, Li Pengpai, Liu Zhiping, Wang Zhigang, Zhang Xiangyan, Ye Xianwei, Handelman Samuel K., Sexton Jonathan, Eaton Kathryn, Higgins Gerry, Omenn Gilbert S., Athey Brian, Smith Barry, Chen Luonan and He Yongqun, "A new framework for host- pathogen interaction research", Frontiers in Immunology, (2022)

[33] Federhen S. (2012) *The NCBI Taxonomy Database*. Nucleic Acids Res. 40:D136-D143. doi:10.1093/nar/gkr1178.

[34] Baltimore D. (1971). *Expression of Animal Virus Genomes*. Bacteriological Reviews. 35, 235-41.

Integrating Declarative and Procedural Knowledge in Infectious Disease Scenario for Epidemiological Monitoring*

Evellin Cardoso[1]

[1]Federal University of Goias, Goias, Brazil

Abstract

Although promising, the synergies between declarative and procedural knowledge have been little explored in Medicine [1]. To tackle this problem, this paper proposes an approach that combines declarative and procedural knowledge. The approach uses ontologies as a knowledge artifact that expresses medical declarative knowledge. This ontology is used as a starting point for an approach that derives a BPMN procedural specification from the knowledge expressed in the ontology. The approach is illustrated in a infectious disease scenario, in particular, it uses the Basic Formal Ontology (BFO) [2] and the Infectious Disease Ontology (IDO) [3].

Keywords

Medical Knowledge, Knowledge Representation, Ontologies, Basic Formal Ontology (BFO), Business Process Management (BPM)

1. Introduction

Expert systems have been used in Medicine successfully since 1990s [4, 5]. An important aspect of the design of such systems regards the acquisition and representation of medical knowledge, a problem that has been addressed by different communities in Computer Science. In medical informatics, a lot of research has been conducted in the development of formalisms to capture medical knowledge. Since late 1990s, many Knowledge Representation (KR) formalisms have been developed, including ontologies, semantic web related formalisms and logics [6]. The focus of medical informatics community is on the representation of (medical) *declarative knowledge* that capture general, background medical knowledge, such as diseases, drugs and treatments [1].

Another strain of research in medical informatics considers the formalization of narrative clinical guidelines into computer interpretable clinical guidelines (CIG) using formalisms such as document models, decision trees, probabilistic models, task-network models [7]. Many domain-specific languages to model CIGs have been proposed, such as Asbru, PROforma, GLIF, EON and GUIDE [8]. More recently, boostered by business demands, Business Process Management (BPM) techniques have been applied to healthcare organizations, in an attempt to streamline

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

*Corresponding author.

evellin@ufg.br,evellinc@gmail.com (E. Cardoso)
0000-0001-6242-662X (E. Cardoso)

medical operations, thus achieving business goals of improving efficiency and reducing costs. Medical processes have been captured as business processes using process languages such as BPMN and DECLARE. Business processes and CIGs capture (medical) *procedural knowledge* that consists of sequences of actions that must be followed by healthcare providers in certain circumstances [1].

Although promising, the synergies between declarative and procedural knowledge have been little explored in both medical informatics and BPM communities [1]. To tackle this problem, this paper proposes an approach that combines declarative and procedural knowledge. The approach uses ontologies as a knowledge artifact that expresses medical declarative knowledge. This ontology is used as a starting point for an approach that derives a BPMN procedural specification from the knowledge expressed in the ontology. The approach is illustrated in a infectious disease scenario, in particular, it uses the Basic Formal Ontology (BFO) [2] and the Infectious Disease Ontology (IDO) [3].

The rest of the paper is structured as follows: Section 2 briefly introduces the BFO and IDO ontologies and procedural models in the medical sector, briefly introducing BPMN syntax. Section 3 presents the approach for deriving a BPMN procedural model starting from the ontologies, illustrating it in the scenario of infectious diseases. Section 5 concludes the paper and outlines future work.

2. Preliminaries

2.1. Ontologies

Ontologies are a knowledge representation formalism that represent (or strive to represent) reality in such way a group of stakeholders understand the terms that compose a certain domain of discourse, and can thus learn about such domain [2]. They may be classified as *top-level ontologies* (or formal) that contain highly general categories and relations of reality common to all domains, defining concepts such as "process", "material object", etc., or *domain ontologies* that capture a basic set of universal concepts pertinent to a specific scientific domain (e.g., geography, medicine or law).

In this paper, (medical) declarative knowledge is represented as ontologies. The *Basic Foundational Ontology (BFO)* [2] is the top-level ontology chose due to its wide accceptance as an international standard ISO/IEC 21838–2, while the Infectious Disease Ontology (IDO) [3] is the chosen domain ontology. IDO extends the Ontology for General Medical Science (OGMS), which in its turn extends BFO.

2.2. Procedural Models in Healthcare

The BPM discipline is concerned with the formalization and analysis of the activities conducted by an enterprise to produce goods and services to its customers [9]. In healthare organizations, medical processes can be divided as *administrative processes* (concerned with administrative practices like handling of medical order and lab procedures) or *knowlege-intensive processes* which are concerned with the intensive usage of domain specific knowledge in diagnose and treament processes.

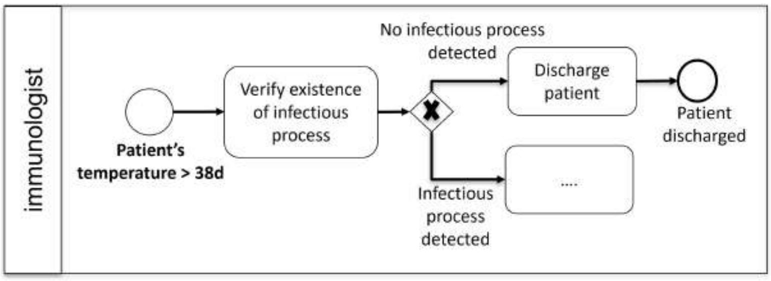

immunologist
Patient's temperature > 38d
Verify existence of infectious process
No infectious process detected
Discharge patient
Patient discharged
....
Infectious process detected

Figure 1: A BPMN specification of knowledge-intensive process in healthcare

Both types of processes can be typically captured by a procedural process language, such as the Business Process Modeling Notation (BPMN). Fig. 1 depicts a knowledge-intensive business process in healthcare. Rounded rectangles denote activities (e.g., *verify the existence of infectious process*), while circles represent events (*patient's temperature > 38 degrees*). Diamonds represent decisions (*presence or absence of infectious process?*), while arrows denote diferent results of decisions.

3. The Approach

This section depicts the approach of building process models based on the knowledge extracted from ontologies. As ontologies are commonly structured into *foundational ontologies* $\mathcal{F}$ and *domain ontology* $\mathcal{D}$, our approach starts from a set of *ontologies* $\mathcal{O} = \langle \mathcal{F}, \mathcal{D} \rangle$, where $\mathcal{F}$ is the set of foundational ontologies and $\mathcal{D}$ the set of domain ontologies. The approach is composed by the following steps:

Step 1. The knowledge engineer starts by investigating the concepts of $\mathcal{O}$ for a preliminary understanding of the domain.

Step 2. As the size of ontologies may be significant, the knowledge engineer selects the subset $\mathcal{D}'$ from $\mathcal{D}$ that is relevant for her modeling purposes. $\mathcal{D}'$ may be sub-ontologies or even portions of $\mathcal{D}$.

Step 3. The knowledge engineer explores the semantics of each particular concept to grasp about the domain.

Step 4. The knowledge engineer builds the procedural representation. To perform such step, s/he considers the following sub-steps (not necessarily in this order):

- One must take into account that activities are performed to react to events happening in the world. The hint is to look for the relevant events $\{e_1, e_2, ..., e_n\}$ taking place within the domain

- To express such events in the procedural specification, consider that the occurrence of events is captured by changes in the states of entities of the domain (concepts of the ontology). By searching for the concepts in the ontology and which changes in the state of these concepts are relevant, one can express the relevant events.

- Order the relevant events. Include the events and the activity that has to be performed to address the event (change), including them all the procedural specification.

Step 5. As procedural and declarative knowledge present a complementary view of reality, it is possible that all knowledge required to build the procedural specification is not found in the declarative specification (and vice-versa). In this case, the knowledge engineer will have to complement the procedural specification with the help of domain experts.

Next section illustrates our approach in a infectious disease scenario.

4. Applying the Approach in an Infection Disease Scenario

This section illustrates the approach from Sec. 3 applied in a infectious disease scenario [3]. The infectious disease scenario represents the domain of infectious diseases. It includes different biological scales (gene, cell, organ, organism and population), complementary disciplinary perspectives (biological, clinical, epidemiological), and successive phases of an infectious process (host, reservoir, vector, pathogen). With the recent breakthrough of Covid-19, the availabillity of such ontology is important to integrate heteregenous data sources. Such integration will enable the establishment public policies to contain the disease by public health organizations using statistical data.

Step 1. Starting the approach, the set of ontogies $\mathcal{O}$ related to the IDO have been investigated for a basic understanding of the infectious disease domain. Three ontologies (BFO, OGMS and IDO) have been identified. In this way we have $\mathcal{O} = \langle \mathcal{BFO}, \mathcal{D} \rangle$, where $\mathcal{D} = \langle \mathcal{IDO}, \mathcal{OGMS} \rangle$.

Step 2. Multiple possibilities for selecting parts $\mathcal{D}'$ of $\mathcal{D}$ exist, given that IDO captures many distinct dimensions (biological scales, disciplinary perspectives and different phases). The disciplinary perspective of *epidemiology* has been chosen.

Step 3. The semantics of each particular concept relative to epidemiology has been investigated at Table 5 and section "Epidemiology and surveillance" from [3].

Step 4. Fig. 2 depicts the BPMN specification built in our approach (the three traces inside the first activity denotes that the activity is performed multiple times, in this case, for each geographic region).

This specification has been built by analyzing the semantics of three IDO concepts (*infectious disease incidence*, *infectious disease pandemic* and *infectious disease epidemic*) that are highlighted in boldface in BPMN. On top of identifying the concepts, we have identified the relevant events.

After understanding the semantics of the three concepts, following with the identification of the events, if the incidence of an infectious disease in a population (in a certain geographic region) is above a certain threshold1 (*infection disease incidence > threshold1*), this may indicate the existence of an *epidemics* in that region. With that, the public administrator has to check the existence of epidemics in other regions as well. If other regions have not surpass the acceptable threshold1 of infectious disease, then the number of regions with infectious disease is below threshold2 (*#regions < threshold2*), no signal of epidemics is found and the surveillance process finishes.

On the contrary, if a number of regions that population has infectious disease surpass a certain threshold (*#regions > threshold2*), this indicates a *pandemics*. In this case, the public administrator proceeds with the monitoring and two situations may happen. One is when

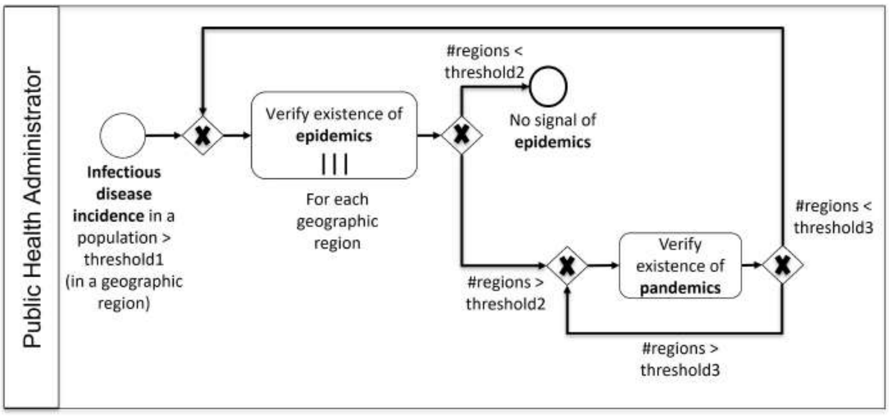

Public Health Administrator
Infectious disease incidence in a population > threshold1 (in a geographic region)
Verify existence of epidemics
For each geographic region
#regions < threshold2
No signal of epidemics
#regions > threshold2
Verify existence of pandemics
#regions < threshold3
#regions > threshold3

Figure 2: Procedural Specification (BPMN) built from an infectious disease ontology [3]

we continuosly have a pandemics (*#regions > threshold3*). The second one happens when the number of regions drop and we no longer have a pandemics, but we may still have an epidemics (*#regions < threshold3*).

Step 5. Notice that the ontology captures only knowledge about the domain, but not the real-time data relative to the incidence of infectious diseases in population or the thresholds for epidemics and pandemics. In this case, public admnistrators have to provide such numbers.

5. Conclusion

This paper proposed a preliminary approach to derive a BPMN procedural specification from the knowledge expressed in an ontology. The approach is illustrated in an infectious disease scenario, using BFO (top-level ontology) and IDO (domain specific ontology). As the integration between declarative and procedural knowledge is generally challenging in many domains (not only Medicine), I hope that this preliminary approach provides an initial insight on how to solve this problem.

In particular, an advantage of this approach in the infectious disease scenario is the possibility of integrating data to the overall approach presented in this paper. Two types of data may be integrated, together with two distinct monitoring perspectives. One perspective considers the usage of the values of incidence of infectious diseases in population and the thresholds for epidemics and pandemics. With these values, it is possible to use epidemic simulators in public health monitoring [10]. The second monitoring approach considers the usage of process mining [9] to monitor the BPMN procedural process, and thus, to monitor in which stage of the process we are (if we have a epidemics, pandemics or no abnormal public health event).

As future work, I consider this integration with data, together with the usage of more concepts from the ontology to derive more information for the procedural specification. Further, I also want to explore the synergy between declarative and procedural knowledge in the other way round, such as, can we derive the ontology from a procedural specification?

References

[1] S. Bragaglia, F. Chesani, P. Mello, M. Montali, Conformance Verification of Clinical Guidelines in Presence of Computerized and Human-Enhanced Processes, Springer International Publishing, 2015, pp. 81–106.

[2] R. Arp, B. Smith, A. D. Spear, Building Ontologies with Basic Formal Ontology, The MIT Press, 2015.

[3] S. Babcock, J. Beverley, L. Cowell, B. Smith, The infectious disease ontology in the age of covid-19, 2021. doi:10.31219/osf.io/az6u5.

[4] P. Windyga, D. Almeida, G. Passariello, F. Mora, J. Coatrieux, Knowledge-based approach to the management of serious arrhythmia in the ccu, Medical and Biological Engineering and Computing 29 (1991) 254–60. doi:10.1007/BF02446707.

[5] A. Batarekh, A. D. Preece, A. Bennett, P. Grogono, Specifying an expert system, Expert Systems with Applications 2 (1991) 285–303. doi:https://doi.org/10.1016/0957-4174(91)90036-E.

[6] D. Riaño, M. Peleg, A. ten Teije, Ten Years of Knowledge Representation for Health Care (2009–2018): Topics, Trends, and Challenges, Artificial Intelligence in Medicine 100 (2019).

[7] M. Peleg, Computer-Interpretable Clinical Guidelines: A Methodological Review, Journal of Biomedical Informatics 46 (2013) 744–763.

[8] D. Isern, A. Moreno, Computer-Based Execution of Clinical Guidelines: A Review, International Journal of Medical Informatics 77 (2008) 787–808.

[9] W. M. P. van der Aalst, Business Process Management: A Comprehensive Survey, ISRN Software Engineering 2013 (2013).

[10] W. Hogan, M. Wagner, M. Brochhausen, J. Levander, S. Brown, N. Millett, J. DePasse, J. Hanna, The apollo structured vocabulary: an owl2 ontology of phenomena in infectious disease epidemiology and population biology for use in epidemic simulation, Journal of Biomedical Semantics 7 (2016) 50. doi:10.1186/s13326-016-0092-y.

Multi-Ontology framework of Maternal Milk for Immune Systems (MOMMIS) - Extended Abstract

Emily Steliotes [1,2], Daniela Barile[1] and Matthew Lange [2]

[1] University of California, Davis, Department of Food Science and Technology, One Shields Ave, Davis, CA, USA
[2] International Center for Food Ontology Operability Data and Semantics (IC-FOODS), 216 F St, Suite #139 Davis, CA, USA

Abstract

We propose a multi-ontology framework modeling the impacts of milk composition on immune health outcomes, using maternal milk as a case study. We evaluated the wide array of immune-health promoting bioactive factors in mammalian milk and existing ontologies relevant to food, nutrition, and health. From here, we are aligning existing ontologies and developing new ontologies to fill notable gaps. MilkOligoDB, which allows for the comparison of milk oligosaccharide profiles among mammalian species and across the literature, demonstrates how the MOMMIS framework can be instantiated. MOMMIS will be useful for interdisciplinary and translational research at the intersection of food and health science disciplines.

Keywords

Maternal Milk, Milk Composition, Immune Systems, Immune Health, Ontologies

1. Immune Health vs. Immune Disease

Neonatal immune cells, with little immunological memory, a developing immune system, and increased vulnerability to a vast array of infectious and non-infectious diseases and conditions, must simultaneously mount responses against environmental stimuli while maturing (1–3). Establishment, development, maturation, optimization and maintenance of immune system functions leads to improved disease resistance, healthspan, and longevity (4,5). Current immune system ontological structures annotate immune system components relative to their corresponding disease states and drug treatments (6,7). Ontological structures are needed, which are capable of annotating immune health system development, maturation, and improvement including both intrinsic and extrinsic factors.

2. Development of Ontologies at the Intersection of Milk Composition and Immune Health

Maternal milk provides the perfect case study to ontologically model food composition impacts on immune health. As mammals' first food, milk confers health from one generation to the next in ways that no other food or biological fluid does. Transmitting evolutionary knowledge via milk, mothers catabolize their own bodies to create and deliver a food providing everything an infant needs, and nothing more. Maternal milk is the only substance consumed through the course of a mammal's lifetime informed by millenia of Darwinian selective pressure to nourish and improve the immune system of its consumer while also setting the stage for improved healthspan and disease resistance (8,9). Children need milk for bone growth, and infants rely on it for crucial immune function as they develop (10,11).

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: esteliotes@ucdavis.edu (Emily Steliotes); dbarile@ucdavis.edu (Daniela Barile); matthew@ic-foods.org (Matthew Lange)
ORCID: 0009-0006-0501-880X (Emily Steliotes); 0000-0002-3889-1596 (Daniela Barile); 0000-0002-6148-7962 (Matthew Lange)

Annotating milk is a worthwhile pursuit because milk plays a crucial role in the growth and development of mammals. An ontologically based maternal milk ⇔ infant health informatics framework offers the opportunity to links bodily tissues and fluids, as well as their packaging and delivery into nutritive foods with the metabolic processes and health outcomes of the consuming organism. Specifically relating to organismal immunological function, ontologically modeling maternal milk and its consumption affords the opportunity to better understand establishment, development, maturation, optimization and maintenance of immune system functions across molecular, cellular, organ, and systems levels. It will also be crucial for interdisciplinary translational research efforts, i.e. determining how an increase in oligosaccharide content in milk affects immune-mediated health outcomes or what immune health-promoting bioactive factor(s) affect the development and/or treatment of emerging viruses such as COVID-19.

2.1. Mammalian Milk Immune Components

Mammalian milk contains a variety of immune-health promoting bioactive factors (milk immune components) including but not limited to: hormones, cytokines, chemokines, lymphocytes, macrophages, neutrophils, T cells, immunoglobulins, lactoferrin, lysozyme, bioactive peptides, antibodies, stem cells, human milk oligosaccharides (HMOs), the microbiota, and microRNAs capable of the mechanisms by which they drive immune maturation (11,12). For example, HMOs play an important role in the prevention of necrotizing enterocolitis, a disease occurring in sick or premature babies (13)(14). *Bifidobacteria infantis*, a probiotic that grows selectively on specific HMOs, is currently used to treat necrotizing enterocolitis (15).

2.1.1. Existing food, nutrition, and health ontologies related to milk immune components

Annotating maternal milk components for their relationships to immune health using ontologies lays the groundwork for a comprehensive food, nutrition, and health informatics framework describing any food components and their corresponding immune health outcomes. Connecting multiple ontologies by building multi-ontology food informatics frameworks allows researchers to ask and answer more interdisciplinary questions (16–19). The Multi-Ontology framework of Maternal Milk for Immune Systems (MOMMIS) connects the following biological ontologies: FoodOn (20), Uberon (21), Compositional Dietary Nutrition Ontology (CDNO) (17), the Mass spectrometry ontology (HUPO) (22), the Mammalian phenotype ontology (23), and the Gene Ontology (GO) (6). Ontologies related to nutrition interventions and personalized nutrition experiments will also be crucial. These include but are not limited to: the Ontology of Precision Medicine and Investigation (OPMI) (24), the Ontology for Biomedical Investigations (OBI) (25), the Ontology for Nutritional Epidemiology (ONE) (26), the Ontology for Nutritional Studies (ONS) (27), the Food Biomarker Ontology (FOBI) (28), the Human Disease Ontology (DO) (29), and the Medical Action Ontology (MaXO) (7), and the Ontology of Host-Microbiome Interactions (OHMI) (30).

2.1.2. Connecting UC Milk and MilkOligoDB: Oligosaccharides as key milk immune components

UC Milk is an ontology that was developed to characterize mammalian milk components and the biological processes giving rise to their creation. The ontology describes both the production and processing of milk itself as well as the role of milk throughout the life cycle, including during three key stages: infant, pregnancy, and lactation (31). MilkOligoDB allows for the comparison of milk oligosaccharide profiles among mammalian species and across the literature. It demonstrates the considerable variation in oligosaccharide profiles both between species and within species (32)**.** Building off of the data model for MilkOligoDB (32), we extend CheBI oligosaccharide classes to cover known classes of mammalian oligosaccharides and then instantiate these classes in the oligosaccharide class (prebiotics) to probiotics in the gut that confer immunological health benefits, including for different types of mammalian milks. Ontologies that allow us to model milk processing conditions such as pasteurization, which are especially relevant for mothers who cannot breastfeed, are being built as well (33).

3. Acknowledgements

We are grateful to the US National Science Foundation, NSF Award numbers RCN 1737573 OAC 2112606, for funding this research and the conceptualization and editing of this Special Collection, both of which were underpinned by our NSF-funded Research Coordination Network, “Developing Informational Infrastructure for Building Smart Regional Foodsheds”.

4. References

[1] Gollwitzer ES, Marsland BJ. Impact of Early-Life Exposures on Immune Maturation and Susceptibility to Disease. Trends Immunol. 2015 Nov;36(11):684–96.

[2] Basha S, Surendran N, Pichichero M. Immune responses in neonates. Expert Rev Clin Immunol. 2014 Sep;10(9):1171–84.

[3] Yu JC, Khodadadi H, Malik A, Davidson B, Salles É da SL, Bhatia J, et al. Innate Immunity of Neonates and Infants. Front Immunol. 2018 Jul 30;9:1759.

[4] Chaplin DD. Overview of the immune response. J Allergy Clin Immunol. 2010 Feb;125(2 Suppl 2):S3–23.

[5] Ji S, Xiong M, Chen H, Liu Y, Zhou L, Hong Y, et al. Cellular rejuvenation: molecular mechanisms and potential therapeutic interventions for diseases. Signal Transduction and Targeted Therapy. 2023 Mar 14;8(1):1–39.

[6] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9.

[7] Carmody LC, Gargano MA, Toro S, Vasilevsky NA, Adam MP, Blau H, et al. The Medical Action Ontology: A Tool for Annotating and Analyzing Treatments and Clinical Management of Human Disease. medRxiv [Internet]. 2023 Jul 13; Available from: http://dx.doi.org/10.1101/2023.07.13.23292612

[8] J. Bruce German FLS. International milk genomics consortium. Trends Food Sci Technol. 2006;(12):656–61.

[9] Sela DA, Mills DA. The marriage of nutrigenomics with the microbiome: the case of infant-associated bifidobacteria and milk. Am J Clin Nutr. 2014 Mar;99(3):697S – 703S.

[10] Carr LE, Virmani MD, Rosa F, Munblit D, Matazel KS, Elolimy AA, et al. Role of Human Milk Bioactives on Infants' Gut and Immune Health. Front Immunol. 2021 Feb 12;12:604080.

[11] Ballard O, Morrow AL. Human milk composition: nutrients and bioactive factors. Pediatr Clin North Am. 2013 Feb;60(1):49–74.

[12] Sampath V, Martinez M, Caplan M, Underwood MA, Cuna A. Necrotizing enterocolitis in premature infants-A defect in the brakes? Evidence from clinical and animal studies. Mucosal Immunol. 2023 Apr;16(2):208–20.

[13] Bode L. Human Milk Oligosaccharides in the Prevention of Necrotizing Enterocolitis: A Journey From in vitro and in vivo Models to Mother-Infant Cohort Studies. Front Pediatr. 2018 Dec 4;6:385.

[14] Tobias J, Olyaei A, Laraway B, Jordan BK, Dickinson SL, Golzarri-Arroyo L, et al. Bifidobacteriumlongum subsp. infantis EVC001 Administration Is Associated with a Significant Reduction in the Incidence of Necrotizing Enterocolitis in Very Low Birth Weight Infants. J Pediatr. 2022 May;244:64–71.e2.

[15] Lange MC, Lemay DG. A multi-ontology framework to guide agriculture and food towards diet and health. of the Science of Food and … [Internet]. 2007; Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/jsfa.2832

[16] Andrés-Hernández L, Blumberg K, Walls RL, Dooley D, Mauleon R, Lange M, et al. Establishing a Common Nutritional Vocabulary - From Food Production to Diet. Front Nutr. 2022 Jun 21;9:928837.

[17] Dooley D, Andrés-Hernández L, Bordea G, Carmody L, Cavalieri D, Chan L, et al. OBO Foundry Food Ontology Interconnectivity [Internet]. [cited 2023 Jul 18]. Available from: https://www.semantic-web-journal.net/system/files/swj3458.pdf

[18] Tomich TP, Hoy C, Dimock MR, Hollander AD, Huber PR, Hyder A, et al. Why Do We Need Food Systems Informatics? Introduction to This Special Collection on Smart and Connected Regional Food Systems. Sustain Sci Pract Policy. 2023 Apr 12;15(8):6556.

[19] Dooley DM, Griffiths EJ, Gosal GS, Buttigieg PL, Hoehndorf R, Lange MC, et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. NPJ Sci Food. 2018 Dec 18;2:23.

[20] Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012 Jan 31;13(1):R5.

[21] Mayer G, Montecchi-Palazzi L, Ovelleiro D, Jones AR, Binz PA, Deutsch EW, et al. The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. Database . 2013 Mar 12;2013:bat009.

[22] Smith CL, Eppig JT. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mamm Genome. 2012 Oct;23(9-10):653–68.

[23] opmi: OPMI: Ontology of Precision Medicine and Investigation [Internet]. Github; [cited 2023 Aug 3]. Available from: https://github.com/OPMI/opmi

[24] Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. PLoS One. 2016 Apr 29;11(4):e0154556.

[25] Yang C, Ambayo H, Baets BD, Kolsteren P, Thanintorn N, Hawwash D, et al. An Ontology to Standardize Research Output of Nutritional Epidemiology: From Paper-Based Standards to Linked Content. Nutrients [Internet]. 2019 Jun 8;11(6). Available from: http://dx.doi.org/10.3390/nu11061300

[26] Vitali F, Lombardo R, Rivero D, Mattivi F, Franceschi P, Bordoni A, et al. ONS: an ontology for a standardized description of interventions and observational studies in nutrition. Genes Nutr. 2018 Apr 30;13:12.

[27] Castellano-Escuder P, González-Domínguez R, Wishart DS, Andrés-Lacueva C, Sánchez-Pla A. FOBI: an ontology to represent food intake data and associate it with metabolomic data. Database [Internet]. 2020 Jan 1;2020. Available from: http://dx.doi.org/10.1093/databa/baaa033

[28] Schriml LM, Munro JB, Schor M, Olley D, McCracken C, Felix V, et al. The Human Disease Ontology 2022 update. Nucleic Acids Res. 2022 Jan 7;50(D1):D1255–61.

[29] He Y, Wang H, Zheng J, Beiting DP, Masci AM, Yu H, et al. OHMI: the ontology of host-microbiome interactions. J Biomed Semantics. 2019 Dec 30;10(1):25.

[30] Colet E, Lange M. uc_Milk: An Ontology for Scientifically-based Unambiguous Characterization of Mammalian Milk, their Composition and the Biological Processes Giving Rise to their Creation. In: ICBO/BioCreative [Internet]. Citeseer; 2016. Available from: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d04a08822664808ed01fe62204ab1b36b7c047fe

[31] Durham SD, Wei Z, Lemay DG, Lange MC, Barile D. Creation of a milk oligosaccharide database, MilkOligoDB, reveals common structural motifs and extensive diversity across mammals. Sci Rep. 2023 Jun 26;13(1):10345.

[32] Salcedo J, Karav S, Le Parc A, Cohen JL, de Moura Bell JMLN, Sun A, et al. Application of industrial treatments to donor human milk: influence of pasteurization treatments, storage temperature, and time on human milk gangliosides. NPJ Sci Food. 2018 Mar 13;2:5.

An Interactive Dashboard for Ontology Quality Monitoring

Avetis Mkrtchian*, Petr Křemen

Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Computer Science

Abstract

Monitoring ontology quality is a key activity during the whole ontology lifecycle that requires adequate tools to provide overview over changing ontologies. In this paper we present an interactive dashboard framework for monitoring ontology metrics and quality indicators. While the framework is designed as generic, we present a prominent use-case for OBO Foundry ontologies, backing by ROBOT metrics and quality indicators.

Keywords

OBO Foundry, ROBOT tool, ontologies, dashboard, OWL

1. Introduction

Ontologies are a well-known paradigm of explicit shared formal conceptualizations. They have been traditionally strongly supported by medicine, or biology [1] that used them as large and rich reference taxonomies. Yet, nowadays, ontologies have become important also for sharing meaning of enterprise and open data, becoming a key piece of data-centric architecture [2].

As a result, communities have been created to monitor and supervise the quality of ontologies of a particular domain, like OBO Foundry [1], Industrial Ontologies Foundry (IOF) [3].

Yet, creating a proper monitoring solution for ontology quality is a challenge. Although most ontologies have been based on the semantic web standards (like RDFS [4], or OWL [5]), their structure can significantly differ, ranging from flat taxonomies to strongly axiomatized logical structures. This complicates creating a reusable solution for monitoring ontology quality, sentencing the communities to develop their own proprietary solutions.

In our work, we offer a framework for building interactive dashboards over a set of ontologies. Our approach to serve different requirements by different communities supervising ontology evolution for monitoring ontology quality and metrics by keeping the generic solution easily and quickly configurable. Having this said, we present here a work in progress. We did a single experiment with the framework so far – creating an interactive dashboard for OBO Foundry ontologies, their quality, metrics and evolution over time.

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

*Corresponding author.

mkrtcave@fel.cvut.cz (A. Mkrtchian); petr.kremen@cvut.cz (P. Křemen)

0009-0003-1678-6773 (A. Mkrtchian); 0000-0001-6299-4766 (P. Křemen)

CEUR Workshop Proceedings (CEUR-WS.org)

Section 2 presents some existing solutions, and in section 3 we present the architecture of our solution together with its features. In section 4 we evaluate our work on a set of OBO Foundry ontologies and discusses its usability with some members of the OBO Foundry community and section 5 presents our conclusions and lessons learnt.

2. Related Work

A traditional view on ontology quality analysis is given in [6]. In [7], an overview of ontology metrics and quality assessment approaches is further elaborated. Various ontology metrics and quality checks can be computed by the ROBOT tool , which is a general-purpose swiss-knife tool for general OWL ontologies, heavily used in the OBO Foundry. This community uses a periodically updated preconfigured dashboard in a tabular form over fundamental quality issues over all OBO Foundry Ontologies [9]. The objective of the OBO Dashboard is to offer a collection of automated tests aimed at defining a baseline level of conformity with OBO Principles and best practices. These principles encompass openness, a common format, URI/identifier space management, versioning, defined scope, textual definitions, relationships, comprehensive documentation, acknowledgment of diverse user communities, clear locus of authority, naming conventions, maintenance, and responsiveness within ontology development and management. However, as stated by OBO Foundry themselves, the outcome of the OBO Foundry Dashboard does not indicate the quality of an ontology's content.

Ontology quality dashboard could be configured also using general-purpose BI solutions over RDF, like SANSA[10].

3. Interactive Dashboard Framework

The Interactive Dashboard Framework is designed to deliver dynamic dashboards over a set of ontologies that can change over time. Its architecture is depicted in Figure 1. One of the key components is ROBOT [8], which provides ability to generate metrics and violation reports for the ontologies. However, the output generated by ROBOT is available in various formats, excluding RDF. To handle RDF data, the RDF4J [11] and Apache Jena [12] libraries are used. The ontological data is stored in the GraphDB [13] database, which communicates with the aforementioned libraries via an API.

For the frontend, the existing dashboard solution Kibana is utilized. Kibana [14] allows visualization, exploration and analysis of data from the Elasticsearch search engine. However, Elasticsearch does not support RDF data, an RDF Indexer is introduced as intermediary between Elasticsearch and GraphDB. The RDF Indexer facilitates the integration of RDF data into Elasticsearch, enabling querying and visualization within the Kibana frontend.

3.1. Main features

One of the main features is the approach to obtaining quality data on ontologies. Since the ROBOT tool is a key component, its capabilities were analyzed in detail, the most suitable commands were: `robot measure` to generate metrics and `robot report` to get a report on

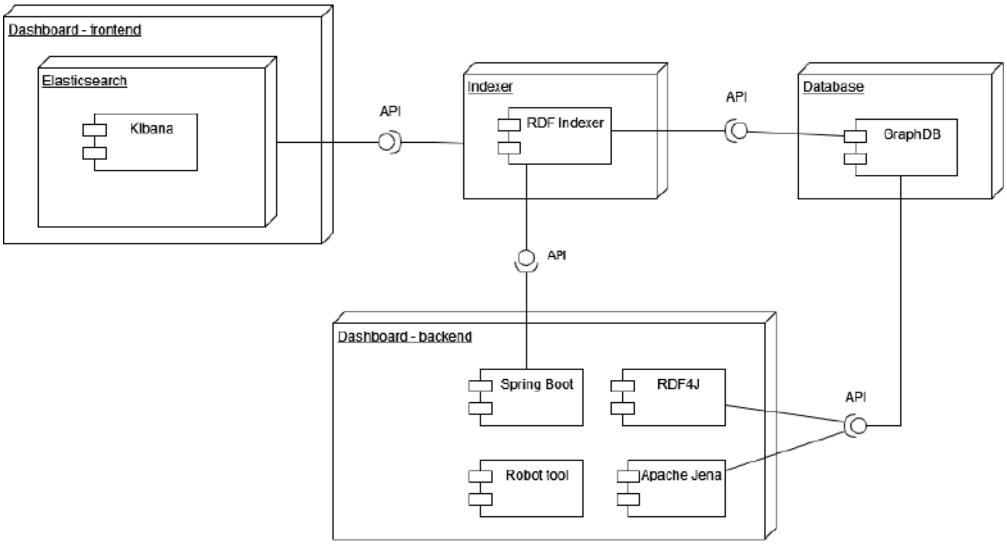


Figure 1: Component diagram

violations in ontologies. It was important for us to represent the outputs of these commands in RDF format in order to save them in triple-store. In order to implement this, the first idea was to describe these two commands using a suitable vocabulary. The results of the research of suitable vocabulary were the finding of DQV(Data Quality Vocabulary)[15] to describe metrics generated by ROBOT. In Figure 2 you can see the part of the data model that is used to represent output metrics of the `robot measure` in RDF format.

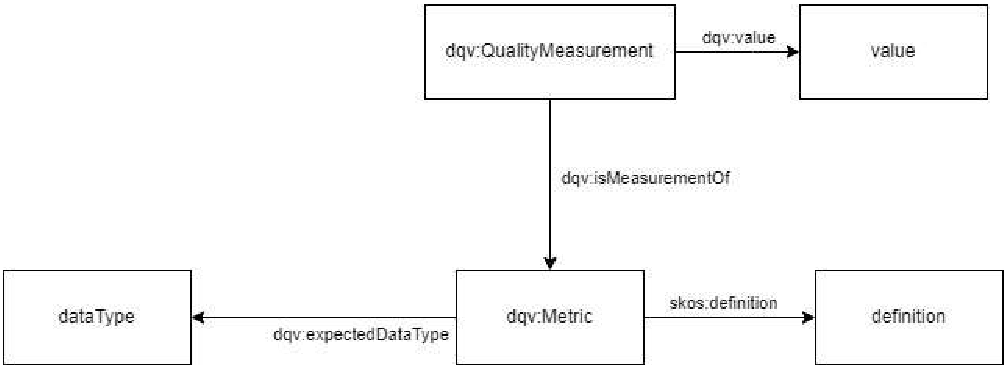


Figure 2: Part of DQV data model

To standardize validation and generation of violation reports and keep them independent of the ROBOT tool we use SHACL (Shapes Constraint Language) [16]. ROBOT offers ontology validation using a set of SPARQL [17] queries which we translated into SHACL.

Here is an example of translating the ROBOT rule called "lowercase definition":

```
PREFIX obo: <http://purl.obolibrary.org/obo/>
```

```
SELECT DISTINCT ?entity ?property ?value WHERE {
  VALUES ?property { obo:IAO_0000115 obo:IAO_0000600 }
  ?entity ?property ?value .
  FILTER (!regex(?value, "^[A-Z0-9]"))
  FILTER (!isBlank(?entity))
}
ORDER BY ?entity
```

This rule can be represented in SHACL as follows:

```
@prefix ex: <http://example.com/ns#> .

ex:lowercase_definition
    a sh:NodeShape ;
    sh:targetClass owl:ObjectProperty , owl:AnnotationProperty ,
    owl:Class ;
    sh:property [
        sh:path obo:IAO_0000115 ;
        sh:severity sh:Info;
        sh:message "lowercase_definition" ;
        sh:pattern "^[A-Z0-9](.*)" ;
    ] ;
    sh:property [
        sh:path obo:IAO_0000600 ;
        sh:severity sh:Info;
        sh:message "lowercase_definition" ;
        sh:pattern "^[A-Z0-9](.*)" ;
    ] .
```

Not all the rules were described exactly as above, complexity of some rules made us replace `sh:property` with `sh:sparql` and adopt the original SPARQL query to its SHACL version. Nevertheless, as many SHACL rules as possible were left without using SPARQL in order to improve validation efficiency. Thus all the predefined queries of the ROBOT report command were transformed in this way, with the exception of two: "deprecated class reference" and "deprecated property reference" due to their complexity.

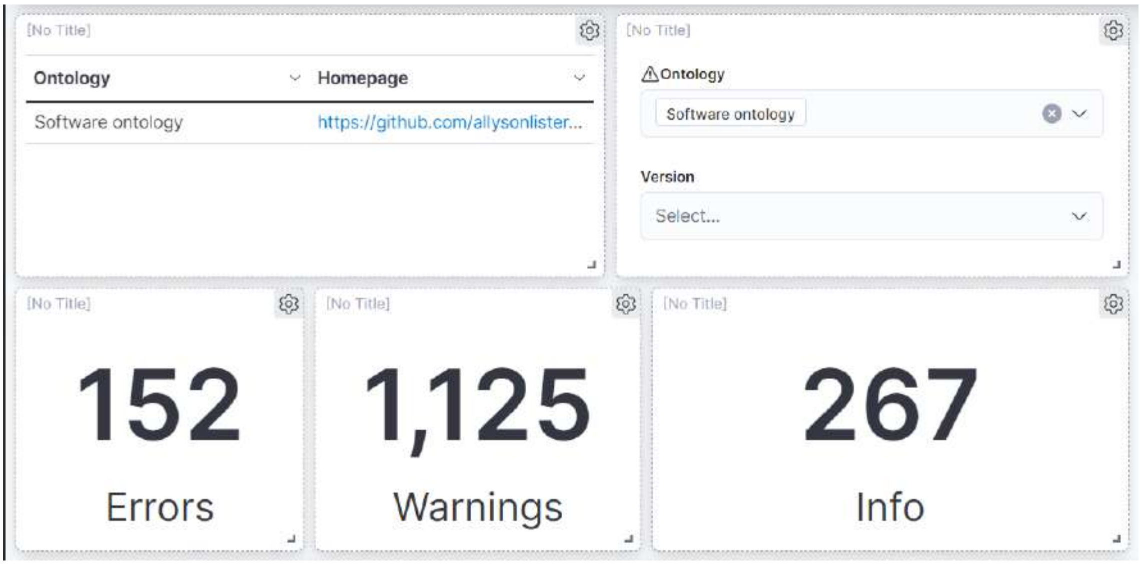
The Interactive Dashboard Framework solves the issue of ontology version tracking, which is currently a significant concern. Many ontology creators either overlook the using of specific OWL attributes like `owl:version` and `owl:versionInfo`, or they they use different schemata as numeric identifiers or date stamps, or a combination of thereof. This inconsistency can lead to difficulties in effectively tracking.

To solve this problem, the Interactive Dashboard Framework works in update mode. With each update, the framework performs a check to determine if the ontology contains date stamp attribute. If such an attribute is absent, the framework assigns its own version to the ontology data, which is date of update. This approach eliminates the reliance on inconsistent versioning practices and ensures ability to track changes over time.

By adopting this strategy, the framework enables the monitoring of variations in the number of violations or specific violations themselves across different ontology versions.

4. Evaluation

In order to test our framework, it was decided to make a prominent use case for OBO Foundry ontologies. The dashboard (available at http://tinyurl.com/obodashboard) currently contains a subset of all OBO Foundry ontologies, mainly to keep the experiment limited and running on a common hardware. While OBO Foundry provides its own dashboard solutions based on its principles, these solutions offer more general information without delving into specifics, such as detailed data on the types of violations. Our dashboard for OBO Foundry ontologies offers easy configuration and possibility to monitor ontology evolution in time. It consists of three main sections: "All ontologies", "Single ontology" and "Specific ontologies". In Figure 3 you can see part of "Single ontology" section, which demonstrates the possibility of selecting an ontology and its version, as well as information about violations of the ontology.



Ontology	Homepage
Software ontology	https://github.com/allysonlister...

Ontology: Software ontology

Version: Select...

152 Errors

1,125 Warnings

267 Info

Figure 3: Selection of one of the ontologies and violations information

Thus, using the example of a dashboard for OBO Foundry ontologies, you can easily configure a dashboard, for this you need to perform only three steps:

1. Configure a SPARQL query tailored to your specific requirements. This query will retrieve the necessary data from GraphDB. By carefully designing your query, you can extract relevant data you wish to visualize.
2. Index data from GraphDB to Elasticsearch with RDF Indexer using the SPARQL query from the previous point.

3. Use Kibana to design customized visualizations, such as graphs, charts, and tables, to effectively represent the ontology data indexed in Elasticsearch.

4.1. User feedback

We received feedback from the OBO Community by conducting four test scenarios that test usefulness where subjects gave comments and rated the scenario on a scale. And also ten questions on usability. Three subjects took part in the tests, and another subject gave his feedback with the format of the discussion.

The test results revealed that most of the comments were focused on the dashboard's UI rather than the data it contains. The primary drawback highlighted by all participants was the height of the dashboard box, which required constant scrolling down to access the most relevant information. This was caused by the navigation bar taking up a large percentage of the available area. Additionally, Kibana has its own UI elements like data filtering fields that further reduce the available space for the dashboard content. The separation of sections in the navigation panel also raised doubts, most participants would like to see the "All ontologies" and "Specific ontologies" sections in one. Almost all participants had problems with filtering data from the first time, this is due to the fact that filtering in Kibana is quite specific and takes a little time to sort it out, as with all other things, since Kibana is a bit technical UI.

As for the positive qualities of the dashboard, the main feature that the community appreciated was the version of ontologies with the possibility of tracking the number of violations in chronological order. One participant was surprised at how we approached the method of generating a violation report, i.e. by transforming ROBOT rules into SHACL syntax, as well as the description of metrics generated by ROBOT using DQV. Another thing appreciated by the participants were the links to the ROBOT rules website and to the ontology repositories.

5. Conclusions

The presented solution shows in a flexible and configurable way to configure an interactive dashboard over ontologies stored in a triple-store (or external files), while still providing useful outputs to the domain experts, as our experiment showed.

In the future, we would like to address several research directions. First, making indexing of large ontologies more efficient and robust. Also, incorporating other types of ontology metrics and quality control rules than those provided by ROBOT. Last, but not least, creating a set of predefined quality control widgets for a default dashboard ready to provide basic quality-related information about any OWL ontology.

This plan for future development will also include integration with the OBO Dashboard, further enhancing its utility and accessibility, as well as trying to apply the dashboard solution to other ontology communities.

References

[1] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. Goldberg, K. Eilbeck, A. Ireland, C. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. Scheuermann, N. Shah, P. Whetzel, S. Lewis, The obo foundry: Coordinated evolution of ontologies to support biomedical data integration, Nature biotechnology 25 (2007) 1251–5. doi:10.1038/nbt1346.

[2] D. McComb, The Data-Centric Revolution: Restoring Sanity to Enterprise Information Systems, Technics Publications, 2019.

[3] Industrial Ontologies Foundry, https://www.industrialontologies.org/, 2023. [Accessed 5-September-2023].

[4] D. Brickley, R. Guha, RDF Schema 1.1, W3C Recommendation, W3C, 2014. Https://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[5] S. Rudolph, P. Patel-Schneider, B. Parsia, M. Krötzsch, P. Hitzler, OWL 2 Web Ontology Language Primer (Second Edition), Technical Report, W3C, 2012. Https://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

[6] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, B. Aleman-Meza, OntoQA: Metric-based ontology quality analysis, in: KADASH, 2005.

[7] R. S. I. Wilson, J. S. Goonetillake, W. A. Indika, A. Ginige, Analysis of ontology quality dimensions, criteria and metrics, in: O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. C. Rocha, E. Tarantino, C. M. Torre (Eds.), ICCSA 2021, Springer International Publishing, Cham, 2021, pp. 320–337.

[8] R. Jackson, J. Balhoff, E. Douglass, N. Harris, C. Mungall, J. Overton, Robot: A tool for automating ontology workflows, BMC Bioinformatics 20 (2019).

[9] OBO Dashboard, https://dashboard.obofoundry.org/dashboard/index.html, 2023. [Accessed 5-September-2023].

[10] J. Lehmann, G. Sejdiu, L. Bühmann, P. Westphal, C. Stadler, I. Ermilov, S. Bin, N. Chakraborty, M. Saleem, A.-C. N. Ngonga, H. Jabeen, Distributed semantic analytics using the sansa stack, in: ISWC'2017, Springer, 2017, pp. 147–155. URL: http://svn.aksw.org/papers/2017/ISWC_SANSA_SoftwareFramework/public.pdf.

[11] E. R. developers, Welcome · Eclipse RDF4J™ | The Eclipse Foundation — rdf4j.org, https://rdf4j.org/, 2023. [Accessed 13-July-2023].

[12] Apache Jena, 2023. URL: https://jena.apache.org/, [Accessed 13-July-2023].

[13] GraphDB Downloads and Resources — graphdb.ontotext.com, 2023. URL: https://graphdb.ontotext.com, [Accessed 13-July-2023].

[14] Kibana: Explore, visualize, Discover Data, 2023. URL: https://www.elastic.co/kibana/, [Accessed 13-July-2023].

[15] R. Albertoni, A. Isaac, Introducing the data quality vocabulary (DQV), Semantic Web 12 (2021) 81–97.

[16] Shapes constraint language (SHACL), Technical Report, W3C, 2017. URL: https://www.w3.org/TR/shacl/, [Accessed 13-July-2023].

[17] E. Prud'hommeaux, A. Seaborne, SPARQL Query Language for RDF, W3C Recommendation, W3C, 2008. URL: https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/, [Accessed 15-July-2023].

Natural Language Processing Tutorial for Biomedical Text Mining - Abstract

Şenay Kafkas[1,2,*], Sumyyah Toonsi[1,2] and Sakhaa Alsaedi[1,2]

[1]*Computer, Electrical and Mathematical Sciences & Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

[2]*Computational Bioscience Research Center, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955, Kingdom of Saudi Arabia*

Abstract

In this tutorial, we introduce Natural Language Processing (NLP) for text mining (TM) in the biomedical domain. The tutorial is structured such that each main concept is backed by hands-on exercises. We start by introducing the difference between NLP and TM. Then continue with motivating the need for text mining in the biomedical domain. Next, we introduce basic concepts of NLP tasks such as Named Entity Recognition (NER), Named Entity Normalization (NEN), and Relationship Extraction (RE. We also cover in detail the widely used methods being used to implement these tasks. These methods include dictionary/ontology-based, rule-based, and advanced machine/deep learning-based approaches. In particular, we cover language models like Word2Vec and transformers (e.g. BERT) and their applications. Furthermore, we discuss the recent advancements in NLP by focusing on the large language models covering GPT, ChatGPT, and others. We conclude our tutorial by discussing limitations and ethics in NLP where we cover the best practices to develop state-of-the-art NLP and TM tools.
The learning objectives of this tutorial are:

- Differences between NLP and TM
- The need for biomedical TM
- Implementation of fundamental NLP tasks: NER, NEN and RE
- Current and future trends in NLP
- Limitations and ethics in NLP and biomedical TM

The learning outcomes of this tutorial are:

- Familiarity with current NLP techniques/tools being used for biomedical TM
- Basic skills to use and develop fundamental NLP tools such as NER and RE
- Familiarity with the current as well as expected future trends in NLP
- Familiarity with the ethics and limitations of biomedical TM

Materials of this tutorial are available from https://github.com/stoonsi/ICBO-NLP-for-Biomedical-Text-Mining-tutorial/tree/main

Keywords

Natural Language Processing, Text mining

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

[*]Corresponding author.

senay.kafkas@kaust.edu.sa (Ş. Kafkas); sumyyah.toonsi@kaust.edu.sa (S. Toonsi); sakhaa.alsaedi@kaust.edu.sa (S. Alsaedi)

0000-0001-7509-5786 (Ş. Kafkas); 0000-0003-4746-4649 (S. Toonsi); 0000-0001-7142-8715 (S. Alsaedi)

CEUR Workshop Proceedings (CEUR-WS.org)

ICBO 2023 Tutorial:
BFO as a top-level ontology for information systems modeling

Mauricio Barcellos Almeida[1] and Jeanne L. Emygdio[2]

[1] *Federal University of Minas Gerais, Belo Horizonte, MG, Brazil*
[2] *Catholic University of Minas Gerais*

Abstract

Abstract text. Basic Formal Ontology (BFO) is a top-level ontology designed to support information retrieval, analysis and integration in scientific and other domains. BFO is used by more than 250 ontology-driven endeavors throughout the world. Indeed, BFO is not only a top-level ontology, rather a framework of resources that provides means to address corporations, including healthcare corporations. Such a framework encompasses middle-level ontologies for several purposes: time and place, information, measures, industry, finance, processes, properties and physical objects. This tutorial explores the potential of BFO for conceptual modeling and it is composed of two parts: The first part aims to introduce basic categories of the world, namely, those adopted within the BFO framework to develop domain ontologies. We will approach the four classification axes from which the primary categories are derived to reach this goal. We will finish the part by presenting the levels of the BFO taxonomy and some theoretical exercises about classification. The second part consists of a hands-on lesson about Protégé, one of the most essential editors in developing ontologies. After some basic notions of the Protégé's UI, we will import BFO and use the editor to classify the entities studied in the first part of the tutorial. We will finish our tutorial with an additional example of corporation modeling.

Keywords

Basic Formal Ontology, top-level ontology, Protégé.

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: mba@eci.ufmg.br (A. 1); jeanne.emygdio@gmail.com (A. 2); ORCID: 0000-0002-4711-270X (A. 1); 0000-0002-7329-4447 (A. 2).

Ontology-Driven Conceptual Modeling with UFO, gUFO, and OntoUML - Abstract

Giancarlo Guizzardi

Semantics, Cybersecurity & Services, University of Twente, Enschede, The Netherlands

Abstract

Conceptual Modeling is about creating concrete artifacts that are meant to represent our conceptualizations of reality for the purpose of communication, domain understanding, problem-solving and meaning negotiation. The artifacts produced by this activity (i.e., conceptual models), thus, serve as an interface between reality and human cognition. For this reason, conceptual modeling languages and conceptual models should be designed by taking very seriously the nature of reality as structured by human cognition, i.e., by systematically employing the-called Descriptive Ontologies. This tutorial revisits a 20-year effort in creating one such Descriptive Ontology, namely, the Unified Foundational Ontology (UFO), as well as a set of conceptual modeling tools based on it. These include the modeling language OntoUML, several patterns and anti-patterns associated with this language, as well as a recent lightweight implementation of UFO termed gUFO (gentle UFO - https://nemo-ufes.github.io/gufo/), which supports the construction of UFO-informed knowledge structuring artifacts (e.g., Knowledge Graphs). The tutorial illustrates these tools in several real-world scenarios.

Keywords

Foundational Ontologies, Conceptual Modeling, UFO, OntoUML, gUFO

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil

EMAIL: g.guizzardi@utwente.nl
ORCID: 0000-0002-3452-553X

Ontology Development in FHIR Resources with the Fast Evidence Interoperability Resources (FEvIR) Platform – Extended Abstract

Brian S. Alper [1,2], Khalid Shahin [1,2], Joanne Dehnbostel [1,2] and Cauê F. Monaco [3]

[1] *Computable Publishing LLC, Ipswich, Massachusetts, United States*
[2] *Scientific Knowledge Accelerator Foundation, Ipswich, Massachusetts, United States*
[3] *Centro Universitário São Camilo, São Paulo, São Paulo, Brasil*

Abstract

Fast Healthcare Interoperability Resources (FHIR) is a standard for health data exchange used globally for electronic health records. Shareable digital objects are called Resources. FHIR defines a CodeSystem Resource structure that can be used to fully represent an ontology.

We extended FHIR to become a standard for data exchange for scientific knowledge. In doing so, we needed to create an ontology, the Scientific Evidence Code System (SEVCO), for expression of study design, risk of bias, and statistics, and we collaborated with the Statistics Ontology (STATO). We created a CodeSystem Resource for the development version of SEVCO (https://fevir.net/27270) and a CodeSystem Resource for the published version of SEVCO (https://fevir.net/sevco).

The Fast Evidence Interoperability Resources (FEvIR) Platform is a free platform supporting the creation, viewing, editing, and storing of scientific knowledge in the form of FHIR Resources. We created FEvIR®: CodeSystem Builder/Viewer to facilitate the creation and viewing of ontologies in the form of FHIR CodeSystem Resources. For the creation of SEVCO, with more than 42 contributors from at least 18 countries, we developed software features and term properties to manage comments, disagreements, agreements, and iterative voting until 100% agreement was reached. We also created FEvIR®: MyBallot to make voting on multiple terms more efficient.

This software demonstration will show the use of FEvIR®: CodeSystem Builder/Viewer and FEvIR®: MyBallot to develop the ontology, and will show the use of FEvIR®: ValueSet Builder/Viewer and Computable Publishing®: Risk of Bias Assessment Tool (RoBAT) as examples of application of the ontology in data curation.

Keywords

Evidence-based medicine, FHIR, data exchange standard, ontology, CEUR-WS

Proceedings of the International Conference on Biomedical Ontologies 2023, August 28th-September 1st, 2023, Brasilia, Brazil
EMAIL: balper@computablepublishing.com (A. 1); kshahin@ computablepublishing.com (A. 2); jdehnbostel@ computablepublishing.com (A. 3), caue.monaco@terra.com.br (A. 4)
ORCID: 0000-0003-4300-4928 (A. 1); 0000-0002-3526-7667 (A. 2); 0000-0002-9896-9842 (A. 3); 0000-0002-5316-2156 (A. 4)