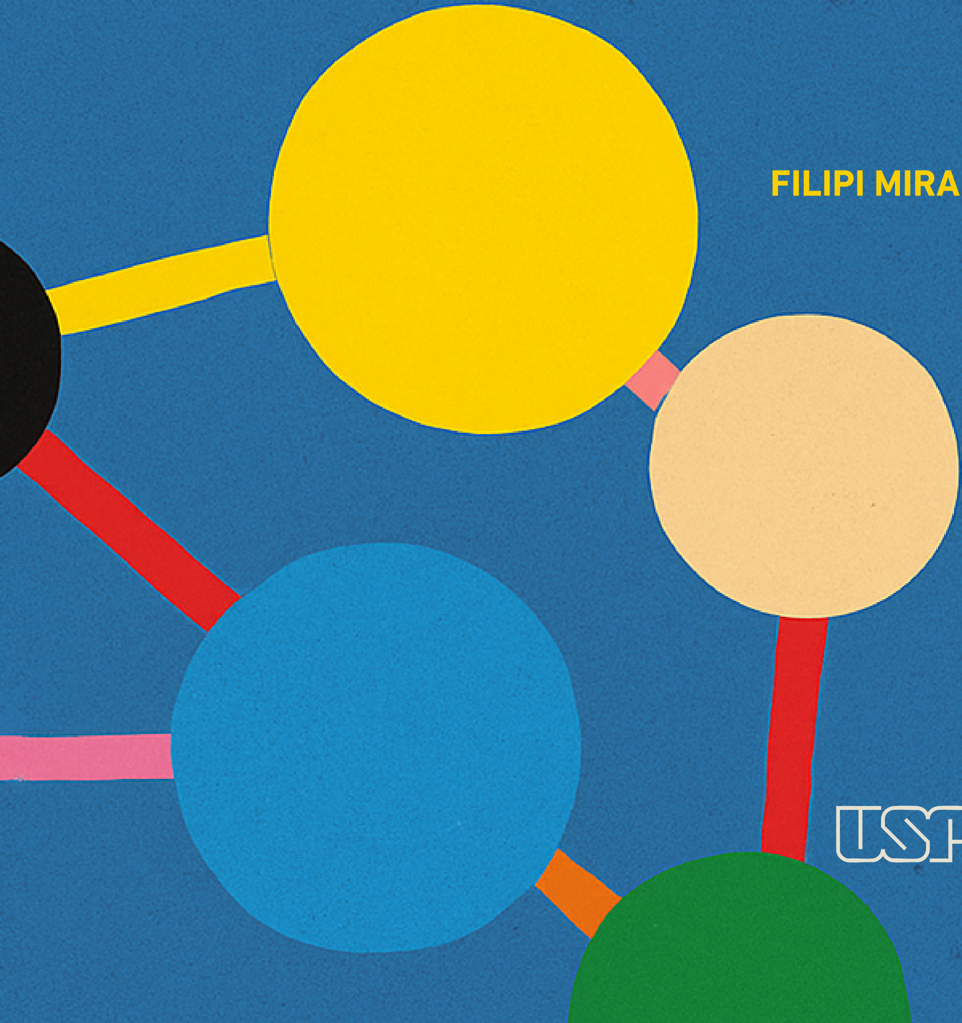
A SEMANTIC INTEROPERABILITY FRAMEWORK FOR DATA-CENTRIC APPLICATIONS IN AGRICULTURE



FILIPI MIRANDA SOARES

USP

UNIVERSITY
OF TWENTE.

A Semantic Interoperability Framework for Data-Centric Applications in Agriculture

Filipi Miranda Soares

A SEMANTIC INTEROPERABILITY FRAMEWORK FOR DATA-CENTRIC APPLICATIONS IN AGRICULTURE

DISSERTATION

to obtain
the degree of doctor at the University of São Paulo,
on the authority of the rector magnificus,
prof.dr. Carlos Gilberto Carlotti Junior,
on account of the decision of the Doctorate Board,
to be defended in a closed examination
on Tuesday, May 27, 2025 at 11.00

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr.ir. A. Veldkamp,
on account of the decision of the Doctorate Board,
to be publicly defended
on Tuesday, October 14, 2025 at 12.45

by

Filipi Miranda Soares
born on January 7, 1996
in Esmeraldas, Minas Gerais, Brazil

This dissertation is approved by:

prof. dr. Antonio Mauro Saraiva (promotor)
dr. Luís Ferreira Pires (promotor)
dr. Luiz Olavo Bonino da Silva Santos (assistant promotor)

University of São Paulo, Polytechnic School,
Department of Computer Engineering and Digital Systems
Av. Prof. Luciano Gualberto, 158 05508-010 Butantã, São Paulo - SP, Brazil

University of Twente, Digital Society Institute,
Semantics, Cybersecurity & Services group
P.O. Box 217, 7500 AE Enschede, The Netherlands

Declaration of AI Use: This dissertation used generative AI tools, specifically OpenAI's ChatGPT, to assist with writing and revision tasks such as paraphrasing, grammar correction, and clarity improvement. All AI-assisted content was reviewed and approved by the author.

Funding: This Ph.D. research was sponsored by the São Paulo Research Foundation (FAPESP), Brazil, under the Grants 21/15125-0 and 22/08385-8.

Cover design: Rubens Rangel Silva
Printed by: Ipskamp Printing
Lay-out: Filipi Miranda Soares
ISBN (print): 978-90-365-6761-9
ISBN (digital): 978-90-365-6762-6
DOI: 10.3990/1.9789036567626

© Filipi Miranda Soares, The Netherlands
All rights reserved. No parts of this dissertation may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author. Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd, in enige vorm of op enige wijze, zonder voorafgaande schriftelijke toestemming van de auteur.

Graduation Committee

At the University of Sao Paulo:

Promotors	prof. dr. A.M. Saraiva	University of Sao Paulo
	dr. L. Ferreira Pires	University of Twente
Co-promotor	dr. L.O.B. da S. Santos	University of Twente
Committee members	dr. G. Sedrakyan	University of Twente
	dr. M. van Keulen	University of Twente
	dr. R. Guizzardi	University of Twente
Reserve members	dr. C.M. Netto	University of Vale do Rio Doce
	prof. dr. F. Cozman	University of São Paulo
	dr. L.S. Maia	Federal University of Ouro Preto
	dr. P.L.P. Correa	University of São Paulo
	dr. R.R. Souza	University of Vienna

At the University of Twente:

Chair/secretary	Prof.dr.ir. B.R.H.M. Haverkort	University of Twente
Promotors	dr. L. Ferreira Pires	University of Twente
	prof. dr. A.M. Saraiva	University of Sao Paulo
Co-promotor	dr. L.O.B. da S. Santos	University of Twente
Committee members	prof. dr. A. Fensel	Wageningen University
	prof. dr. F. Cozman	University of São Paulo
	prof. dr. J. van Hillegersberg	University of Twente
	dr. M. van Keulen	University of Twente
	dr. R. Guizzardi	University of Twente

Acknowledgments

From an early age, I aspired to work in science – a journey that has been both challenging and rewarding. Every step forward was made possible by the generous, wise, and kind-hearted people I met along the way, to whom I am deeply grateful.

I was fortunate to begin my research journey in the second year of my undergraduate studies. Professor Gercina saw potential in me and introduced me to Benildes, who became a mentor and guide in the world of research, and later my master's supervisor. Her unwavering support – at times extending beyond academic matters – helped me publish my first paper in 2015. I am grateful that our partnership and friendship have endured to this day. Thank you so much for everything.

During my master's, I had the privilege of working with Debora, whose expertise in agricultural biodiversity and metadata standards opened many doors for me. More importantly, I gained a dear friend. It was through Debora that I attended my first international conference, BiodiversityNext, in Leiden – an experience that ultimately brought me back to the Netherlands years later. It was also at that event that I met Professor Saraiva, who would become my PhD supervisor at the University of São Paulo (USP).

Professor Saraiva's kindness and his passion for science that drives positive social change have been truly inspiring. Beyond the scope of my PhD research, he encouraged me to engage in initiatives such as the Planetary Health Alliance, SURPASS, Pomar Urbano, etc., which profoundly reshaped my perspective on environmental responsibility. I am especially grateful for his guidance, patience, constant encouragement, and for the many opportunities he opened for me along the way.

In 2021, during the pandemic, a conversation with Professor Luiz Olavo at a virtual conference introduced me to the University of Twente (UT) and shifted the course of my academic path. He became my daily supervisor and introduced me to my promotor, Professor Luís Pires. At UT, I had the opportunity to delve deeply into FAIR data principles and the semantic web – far more than I could have imagined. I'm especially thankful to Professor Pires for always supporting my ideas and providing detailed feedback on every text we wrote together. His constructive input has made me grow immensely. To Professor Bonino, I'm also deeply grateful for guiding me through the complexities of FAIR – his expertise and mentorship were instrumental to my learning.

To my family, who always believed in my dream of pursuing a career in science – even without knowing how I could overcome the many obstacles, especially the financial ones, to make it to college, they never doubted me. Their faith gave me the strength to keep going. To my father, Taquinho; my mother, Néia; my sisters, Tha and Vi; my grandmothers, Elsa and Divina; and my grandfathers, Ladin and Luís (in memoriam) – thank you for your unconditional love and support. And to the newest member of our family, who was born just over a year ago and has already brought so much joy into my life – my nephew and godson, João Miguel.

To my life partner, Rubens, who also became my partner in research – thank you for always believing in me, for standing by my side through everything, and for understanding my need to be away while pursuing my PhD.

To my dear friends Ana, Andrea, Daniel, Fran, Marília, Moritz, Pablo – thank you for keeping our bond strong, no matter the distance.

To my colleagues at C4AI-USP – Alexandre, Ana, Bruno, Celso, Cláudio, Debora, Dilvan, Eliana, Emilian, Enio, Erica, Fábio, Fernando Correa, Fernando Osório, Fray, Jairo, Kelly, Laurence, Marlon, Saraiva, and Wilian – thank you for creating such a supportive and inspiring environment throughout this journey.

To everyone at the UT SCS group – Andrea, Azqa, Barbara, Bertine, Chakshu, Claudenir, Federico, Florian, Geert Jan, Giancarlo, Glenda, Ítalo, Isadora, Jerre, João, Jorik, Laura, Luca, Luís, Luiz, Maarten, Marc, Marten, Matteo, Mattia, Pedro, Pedro Paulo, Rodrigo, Taru, Thijs, Tiago, Willem, and Zsolt – thank you for the enriching discussions, friendship, and constant support.

To the Department of Computer and Digital Systems Engineering (PCS-USP), for the continuous support and everything I learned from its professors and staff – especially Anarosa, Anna Helena, Bruno, Cíntia, Edson, Jorge Rady, Mariza, and Pedro. I also extend my thanks to everyone at the Graduate Program in Electrical Engineering Secretariat for the guidance and assistance throughout my PhD.

To Donika – thank you for the wonderful friendship we built and the countless joyful moments we shared. You brought so much light into my life in the Netherlands. I'm also grateful to you for introducing me to the kind-hearted Teun, who has since become a lifelong friend.

Isadora, Renata, Rodrigo, and Vitor – you became my family away from home and made Brazil feel closer. I'm truly grateful for your friendship.

A special thank you to my *bicho-grilo* friends from the SURPASS2 team: Betina, Bruno, Debora, Kayna, Natalia, Paty, Paula, Salim, Sarinha, Sheina, and Tiago.

To CODATA, RDA, and the Cross-Domain Integration (CDI) Working Group for all the exchanges, including two Dagstuhl workshops. These experiences broadened my understanding of metadata and the semantic web, and introduced me to brilliant colleagues: Arofan, Benjamin, Dan, Darren, Deirdre, Flavio, Hilde, Jay, Joachim, John, Larry, Noemi, Oliver, Pierre-Antoine, Simon, Steven, and Wendy.

To my colleagues from Planetary Health Brazil and the National Institute on Research and Technology on Combating Hunger and Undernutrition – Aline, Daniela, Patricia, Raquel, and Semíramis – thank you for the meaningful work we have done together.

My appreciation also goes to the GoFAIR Brazil Agriculture team: Ana Paula, Benildes, Bibiana, Debora, Ivan, Marcia, Maria Carolina, Maria de Cléofas, Milena, Sérgio Manuel, and Victor Paulo. To the WorldFAIR project team – Debora, Jorit, Salim, Simon, and Arofan – for the collaboration and shared learning.

To my colleagues who worked alongside me on the SmartFood workshop – Renata, Gayane, and Cath – thank you for your collaboration.

To the Pomar Urbano team – Aline, Antonio, Ana Carolina, Anand, Bárbara, Benildes, Debora, Hillary, Isabela, Karoline, Lidio, Luís, Maria Carolina, Maria Clara, Natalia, Raquel, Rubens, Sheina, Stephanie, Uiara, Yamine, and Ylana – thank you for turning a long-standing dream into a tangible and impactful project.

To my defense committee, for their invaluable feedback: Anna, Cristiane, Fabio, Gayane, Jos, Lucinéia, Maurice, Pedro, Renata, and Renato.

To the São Paulo Research Foundation (FAPESP), for the grants n. 21/15125-0 and 22/08385-8, which made this double degree possible.

Finally, my heartfelt thanks to all those who, in one way or another, have been part of this journey. Your support, guidance, and friendship made all the difference.

Filipi M. Soares,
Enschede, October 14, 2025

A Perfeição	**Perfection**
O que me tranquiliza é que tudo o que existe, existe com uma precisão absoluta.	What reassures me is that everything that exists exists with absolute precision.
O que for do tamanho de uma cabeça de alfinete não transborda nem uma fração de milímetro além do tamanho de uma cabeça de alfinete.	What is the size of a pinhead does not overflow by even a fraction of a millimeter beyond the size of a pinhead.
Tudo o que existe é de uma grande exatidão. Pena é que maior parte do que existe com essa exatidão nos é tecnicamente invisível.	Everything that exists is of great exactness. The pity is that most of what exists with such exactness is technically invisible to us.
O bom é que a verdade chega a nós como um sentido secreto das coisas.	The good thing is that truth reaches us as a secret meaning of things.
Nós terminamos adivinhando, confusos, a perfeição.	We end up guessing, confused, at perfection.

Clarice Lispector. *A Descoberta do Mundo*. Rio de Janeiro: Rocco, 1984.

Abstract

The rapid growth of data-centric applications in agriculture has generated vast and heterogeneous datasets, yet their potential is constrained by the lack of semantic interoperability, which limits meaningful data exchange, integration, and reuse. This dissertation proposes a Semantic Interoperability Framework that integrates metadata schemas, ontologies, knowledge graphs, and artificial intelligence to resolve interoperability conflicts in naming conventions, domain representation, and metadata alignment, while adhering to the FAIR data principles. Developed through a Design Science Research methodology, the framework combines structured metadata annotation, ontological modeling for semantic alignment, and knowledge graph construction to enhance data linking and reasoning, with a large language model (LLM) supporting knowledge graph generation and creating SPARQL queries from natural language prompts. Its applicability is demonstrated through two case studies: (1) Agricultural Price Index Data in Brazil, which aligns datasets from CEPEA, IPEA, and CONAB using the Almes Core metadata schema and the APTO ontology for agricultural product types; and (2) Agrobiodiversity and Plant–Pollinator Interaction Data within the WorldFAIR project, which shows how FAIR-aligned schemas and ontology-driven integration standardize complex ecological datasets for scientific collaboration. Evaluation through ontology validation metrics, usability testing, and query performance demonstrates significant improvements in data interoperability, enabling more accurate retrieval, integration, and machine-driven reasoning. The findings also highlight persistent challenges such as metadata adoption, automation of ontology construction, and the need for stakeholder engagement in standardization efforts. Overall, this research offers a novel, scalable, and reusable approach to achieving semantic interoperability in agriculture, bridging fragmented datasets and advancing open data initiatives, digital agriculture policies, and AI-driven analytics.

Keywords: Artificial Intelligence, Semantic Web, Linked Data, Knowledge Graph, Agriculture, Semantic Interoperability, Large Language Model, Agricultural Economics, Agricultural Biodiversity, FAIR Data Principles.

Resumo

O rápido crescimento das aplicações centradas em dados na agricultura tem gerado conjuntos de dados vastos e heterogêneos, mas seu potencial é limitado pela falta de interoperabilidade semântica, que restringe a troca, integração e reutilização significativa das informações. Esta tese propõe um Framework de Interoperabilidade Semântica que integra esquemas de metadados, ontologias, grafos de conhecimento e inteligência artificial para resolver conflitos de interoperabilidade em convenções de nomenclatura, representação de domínio e alinhamento de metadados, em conformidade com os princípios FAIR. Desenvolvido por meio da metodologia de Design Science Research, o framework combina anotações estruturadas de metadados, modelagem ontológica para alinhamento semântico e construção de grafos de conhecimento para aprimorar o vínculo e o raciocínio sobre os dados, com o apoio de um modelo de linguagem de grande porte (LLM) na geração de grafos de conhecimento e na criação de consultas SPARQL a partir de comandos em linguagem natural. Sua aplicabilidade é demonstrada em dois estudos de caso: (1) Dados de Índices de Preços Agrícolas no Brasil, que alinha conjuntos de dados do CEPEA, IPEA e CONAB por meio do esquema de metadados Almes Core e da ontologia APTO para padronização de tipos de produtos agrícolas; e (2) Dados de Agrobiodiversidade e Interação Planta–Polinizador no âmbito do projeto WorldFAIR, que evidencia como esquemas alinhados aos princípios FAIR e abordagens baseadas em ontologias podem padronizar e integrar conjuntos de dados ecológicos complexos, promovendo a colaboração científica. A avaliação por métricas de validação ontológica, testes de usabilidade e desempenho de consultas demonstra melhorias significativas na interoperabilidade de dados, permitindo recuperação mais precisa, integração e raciocínio automatizado por meio de representações em grafos de conhecimento. Os resultados também evidenciam desafios persistentes, como a adoção de metadados, a automação da construção de ontologias e a necessidade de engajamento de stakeholders nos esforços de padronização. Em síntese, esta pesquisa oferece uma abordagem inédita, escalável e reutilizável para alcançar a interoperabilidade semântica na agricultura, conectando conjuntos de dados fragmentados e impulsionando iniciativas de dados abertos, políticas de agricultura digital e análises orientadas por inteligência artificial.

Palavras-chave: Inteligência Artificial, Web Semântica, Dados Ligados, Grafo de Conhecimento, Agricultura, Interoperabilidade Semântica, Modelo de Linguagem de Grande Escala, Economia Agrícola, Biodiversidade Agrícola, Princípios FAIR de Dados.

Samenvatting

De snelle groei van datacentriche toepassingen in de landbouw heeft geleid tot omvangrijke en heterogene datasets, maar het potentieel ervan wordt beperkt door het gebrek aan semantische interoperabiliteit, waardoor betekenisvolle uitwisseling, integratie en hergebruik van gegevens wordt bemoeilijkt. Dit proefschrift stelt een Semantisch Interoperabiliteitsraamwerk voor dat metadata-schema's, ontologieën, kennisgrafen en kunstmatige intelligentie integreert om interoperabiliteitsconflicten op te lossen met betrekking tot naamgevingsconventies, domeinrepresentatie en metadata-afstemming, in overeenstemming met de FAIR-dataprincipes. Ontwikkeld volgens de Design Science Research-methodologie combineert het raamwerk gestructureerde metadata-annotatie, ontologische modellering voor semantische afstemming en de constructie van kennisgrafen om gegevenskoppeling en redeneren te verbeteren, met ondersteuning van een groot taalmodel (LLM) voor de generatie van kennisgrafen en het opstellen van SPARQL-query's vanuit natuurlijke taal. De toepasbaarheid wordt aangetoond aan de hand van twee casestudy's: (1) Landbouwprijsindexgegevens in Brazilië, waarbij datasets van CEPEA, IPEA en CONAB worden geharmoniseerd met behulp van het metadata-schema Almes Core en de APTO-ontologie voor landbouwproducttypen; en (2) Agrobiodiversiteit en plant–bestuiver interacties binnen het WorldFAIR-project, dat laat zien hoe FAIR-gebaseerde schema's en ontologiegestuurde integratie complexe ecologische datasets standaardiseren en integreren om wetenschappelijke samenwerking te bevorderen. Evaluatie met behulp van ontologievalidatiemaatstaven, bruikbaarheidstesten en queryprestaties toont aanzienlijke verbeteringen in gegevensinteroperabiliteit, wat leidt tot nauwkeurigere terugwinning, integratie en machinegestuurde redenering via kennisgraafrepresentaties. De bevindingen benadrukken ook blijvende uitdagingen, zoals de adoptie van metadata, de automatisering van ontologieconstructie en de noodzaak van betrokkenheid van belanghebbenden bij standaardisatie-inspanningen. Al met al biedt dit onderzoek een nieuwe, schaalbare en herbruikbare aanpak voor het realiseren van semantische interoperabiliteit in de landbouw, door gefragmenteerde datasets met elkaar te verbinden en open data-initiatieven, digitaal landbouwbeleid en AI-gedreven analyses te versterken.

Trefwoorden: Kunstmatige Intelligentie, Semantisch Web, Gekoppelde Gegevens, Kennisgrafiek, Landbouw, Semantische Interoperabiliteit, Groot Taalmodel, Landbouweconomie, Agrarische Biodiversiteit, FAIR-gegevensprincipes.

Contents

List of Figures xi

List of Tables xv

List of Acronyms xvii

I Setting the Stage 1

1 Introduction 5
- 1.1 Background 5
- 1.2 Interoperability 8
- 1.3 Semantic Interoperability 9
- 1.4 Related Work on Semantic Interoperability 14
- 1.5 Semantic Interoperability in Agriculture 16
- 1.6 Hypothesis 18
- 1.7 Objective 18
- 1.8 Research Design 19
- 1.9 Dissertation Structure 22

2 Framework for Semantic Interoperability 25
- 2.1 Introduction 25
- 2.2 Framework Design 26
- 2.3 Goal Definition 27
- 2.4 Data Preparation 28
- 2.5 Data Semantic Annotation 29
- 2.6 Validation 33

3 A Conceptual Model for FAIR Metadata Schemas 35
- 3.1 Introduction 35
- 3.2 Related Work 36
- 3.3 A Brief Introduction to OntoUML Stereotypes 36
- 3.4 Conceptual Model 39
- 3.5 Final Considerations 47

II Case Study I: Harmonization and Integration of Agricultural Price Index Datasets 49

4 Goal Definition and Data Preparation 53
4.1 Goal Definition 53
4.2 Selected Datasets 55
4.3 Creating the Data Template 59
4.4 Data Preparation 60

5 Semantic Annotation 63
5.1 Searching for Metadata Schemas 63
5.2 Creating a New Metadata Schema 64
5.3 Mapping the Datasets to the Almes Core Schema 73

6 The Agricultural Product Types Ontology 77
6.1 Classification Systems on Agricultural Product Types . . . 77
6.2 Ontologies for Agricultural Product Types 81
6.3 Development Process 83
6.4 Purpose Identification and Requirements Elicitation 84
6.5 Ontology Capture and Formalization 85
6.6 Ontology Design and Implementation in OWL 88

7 Ontology Design Supported by an LLM 93
7.1 Introduction 93
7.2 Related Work 95
7.3 Methods 96
7.4 Approach 1 101
7.5 Approach 2 111
7.6 Integrating the `:Organism` and `:Product_type` Modules . 117
7.7 Final Considerations 122

8 Ontology Testing and Validation 125
8.1 Introduction 125
8.2 Related Work 126
8.3 User Experience (UX) 127
8.4 APTO Evaluation Using Tree Testing 129
8.5 Results 133
8.6 Comparative Analysis of Task Results 149
8.7 Detailed Analysis of Results 152
8.8 Discussion 164
8.9 FAIR Assessment 167
8.10 Ontology Publication 170
8.11 Final Considerations 170

9 Knowledge Graph for Agricultural Price Data 171
9.1 Knowledge Graph Design Methodology 171

9.2 Use and Applicability of Price Index Data 174
9.3 Knowledge Graph Structure 174
9.4 Creating the Metadata Graph 176
9.5 Extracting GeoNames Data for Locations 177
9.6 Creating the Data Graph 179
9.7 Future Work 181

10 Querying and Validating the Knowledge Graph Using LLM 183
10.1 Implementation of a Triple Store 183
10.2 Generating Queries with an LLM 184
10.3 Query Examples 185
10.4 Benefits and Limitations 197
10.5 FAIR Assessment 200
10.6 Final Considerations 204

III Case Study II: Harmonization and Integration of Plant-Pollinator Interactions Datasets 207

11 Semantic Interoperability in The WorldFAIR Project 211
11.1 Introduction 211
11.2 Goal Definition 212
11.3 Data Preparation 213
11.4 Data Semantic Annotation 214
11.5 Validation 218
11.6 Comparative Analysis of Case Studies 1 and 2 223
11.7 Final Considerations 226

12 Final Remarks 229
12.1 Addressing the Semantic Interoperability Issues 230
12.2 Research Objectives and Outcomes 231
12.3 Contributions 232
12.4 Generalizability of Findings 234
12.5 Limitations 234
12.6 Future Work 236

Data and Code Availability 239

Author Publications 243

Bibliography 253

List of Figures

1.1 The Broken Telephone Game Metaphor. 10
1.2 Information System Integration Using Ontologies 14
1.3 Example of Semantic Incompatibilities within Datasets. 17
1.4 The Elements of an Artifact 20

2.1 Goal Definition Process 27
2.2 Data Preparation Process 29
2.3 Data Semantic Annotation Process 30
2.4 Creating Metadata Schema Sub-Process. 31
2.5 Creating Controlled Vocabulary as Ontology Sub-Process . . . 32
2.6 Validation Process. 34

3.1 Conceptual Model of a FAIR Metadata Schema. 40

5.1 Extended Almes Core Conceptual Model 66

6.1 Organism Module in OntoUML. 86
6.2 Product Type Module in OntoUML. 87
6.3 APTO Product Type Module in OWL. 90
6.4 Synonym Indicated as Preferred Term in Agrotermos 91

7.1 Workflow for the automated ontology design. 98
7.2 Duplicated Labels and Relationships in the Ontology Visualization in Protégé. 110
7.3 Classes Hierarchy on Protégé 114
7.4 ‘Pimenta’ and its Translation as Black Pepper in Wikipedia. . 119
7.5 Herbs Included in ‘cheiro verde’ by Brazilian Regions. 120

8.1 APTO Prototype Upper Classes. 133
8.2 Task 1 - Whey - Scores. 136
8.3 Task 2 - Soybean Meal - Scores. 138
8.4 Task 3 - Boi Magro, Boi Gordo, Vaca Gorda, Vaca Leiteira - Scores. 139

8.5 Task 4 - Frozen Chicken, Chilled Chicken - Scores. 140
8.6 Task 5 - Dairy Drink, Yogurt, Powdered Milk, Pasteurized Milk, UHT Milk, Cheese - Scores. 142
8.7 Task 6 - Araucaria, Eucalyptus, Pinus - Scores. 143
8.8 Task 7 - Andiroba almond, Cocoa bean, Cashew nut, Coconut - Scores. 145
8.9 Task 8 - Pineapple pulp, Açaí pulp, Buriti pulp, Ceriguela pulp, Cupuaçu pulp - Scores. 146
8.10 Task 9 - Rice flakes, Corn flakes - Scores. 147
8.11 Task 10 - Cane sugar and Ethanol - Scores. 148
8.12 Task 11 - Babassu oil, Macauba oil, Buriti oil, Copaiba oil, Murumuru oil, Pequi oil - Scores. 150
8.13 Pietree visualization for Task 1 - Classification of Whey . . . 153
8.14 Pietree visualization for Task 2 – Classification of Soybean Meal and Wheat Bran 154
8.15 Pietree visualization for Task 3 – Classification of Boi Magro, Boi Gordo, Vaca Gorda, and Vaca Leiteira 155
8.16 Pietree visualization for Task 4 – Classification of Frozen Chicken and Chilled Chicken 156
8.17 Pietree visualization for Task 5 – Classification of Dairy Drink, Yogurt, Powdered Milk, Pasteurized Milk, UHT Milk, and Cheese 157
8.18 Pietree visualization for Task 6 – Classification of Araucaria, Eucalyptus, and Pinus 158
8.19 Pietree visualization for Task 7 – Classification of Andiroba almond, Cocoa bean, Cashew nut, and Coconut. 159
8.20 Pietree visualization for Task 8 – Classification of Pineapple pulp, Açaí pulp, Buriti pulp, Ceriguela pulp, and Cupuaçu pulp 160
8.21 Pietree visualization for Task 9 – Classification of Rice flakes and Corn flakes 162
8.22 Pietree visualization for Task 10 – Classification of Cane sugar and Ethanol 163
8.23 Pietree visualization for Task 11 – Classification of Babassu oil, Macauba oil, Buriti oil, Copaiba oil, Murumuru oil, and Pequi oil 164
8.24 APTO FAIR Assessment on Agroportal. 167

9.1 C4AI-KGAP Conceptual Model. 175
9.2 Metadata Converter Pipeline. 177
9.3 Geonames Converter Pipeline. 178
9.4 Data Converter Pipeline. 180

10.1 Query Results - Mean Price for 'Boi Gordo' in 2022. 189
10.2 Query Results - Prices from CEPEA and CONAB for Cane Sugar for the First Semester of 2023. 193

10.3 Harmonized query results showing prices from CEPEA and CONAB for cane sugar for the first semester of 2023. 195
10.4 Prices for Soybean in Paraná and Mato Grosso in 2023. 197
10.5 Assessment Results for the Cepea Dataset. 202
10.6 Assessment Results for the Ipea Dataset. 202
10.7 Assessment Results for the CONAB Dataset. 203
10.8 Assessment Results for the C4AI-KGAP. 203

11.1 Overview of the Data Schema for Plant–Pollinator Interactions 216
11.2 Access Levels for WorldFAIR Pilots on the GloBI Webpage . . 218
11.3 Query and Search Results for *Lasius alienus* visiting *Trinia glauca* 219
11.4 Example of the First Page of a Data Review Generated by GloBI Bots 221

List of Tables

3.1 Adaptation of the FAIR Data Principles for Metadata Schema Management . . . 38
3.2 Attributes for a Schema's Metadata Record . . . 42

4.1 Snippet of the CEPEA Database in Portuguese . . . 56
4.2 Snippet of the IpeaData Database in Portuguese . . . 57
4.3 CONAB Weekly Prices for Fed Cattle in Portuguese . . . 58
4.4 CONAB Monthly Prices for Soybean in Portuguese . . . 59

5.1 Namespaces of Metadata Schemas and Vocabulary Encoding Schemes Reused by Almes Core . . . 68
5.2 Schema Elements with Domain and Range. . . . 69
5.3 Template for Calculating the FAIRness Level. . . . 72
5.4 Mapping of Elements from the Metadata Template to Extended Almes Core. . . . 74
5.5 Mapping of Elements from the Data Template to Almes Core, with a Single Date as Reference for the Period. . . . 74
5.6 Mapping of Elements from the Data Template to Almes Core, with a Range of Dates as Reference for the Period. . . . 74

6.1 APTO Namespace Prefixes and IRIs . . . 89

7.1 Relationship Between Related Studies and this Chapter . . . 97
7.2 Prompt Specification . . . 102
7.3 List of Animal Scientific Names . . . 102
7.4 Deprecated Scientific Names (Synonyms) and Corresponding Accepted Names Detected by ChatGPT. . . . 106
7.5 Iteration with ChatGPT . . . 108
7.6 Iteration with ChatGPT: Fixing Repetitions . . . 111
7.7 Iteration with ChatGPT: Fixing Relationships . . . 112
7.8 Classification of *Piper nigrum* and *Capsicum* sp. . . . 118

8.1 Participant Task Performance Summary . . . 134
8.2 Task Performance Summary . . . 149

10.1 Price of Coffee Beans in Minas Gerais on June 1, 2023 187
11.1 Comparison of Case Studies 1 and 2 223

List of Acronyms

AI	Artificial Intelligence
Almes Core	Agriculture and Livestock Metadata Elements Set
APTO	Agricultural Products Type Ontology
AVM	Astronomy Visualization Metadata
BCO	Biological Collections Ontology
BINs	Barcode Index Numbers
BPMN	Business Process Model and Notation
C3PO	Crop Planning and Production Process Ontology
C4AI	Center for Artificial Intelligence
KGAP	Knowledge Graph on Agricultural Prices
CDI	Cross Domain Integration
CDNO	Compositional Dietary Nutrition Ontology
CEPEA	Center for Advanced Studies on Applied Economics
CGU	Comptroller General of the Union
CONAB	National Supply Company
DSW	Darwin-SW
DC	Dublin Core
DCAT	Data Catalog Vocabulary
DCMI-COLLFREQ	Collection Description Frequency Vocabulary
DDC	Dewey Decimal Classification
DDI	Data Documentation Initiative
DoPI	Database of Pollinator Interactions
DwC	Darwin Core
DwC-A	Darwin Core Archive
EER	Enhanced Entity–Relationship

EHR	Electronic Health Records
EOL	Encyclopedia of Life
EML	Ecological Metadata Language
Embrapa	Brazilian Agricultural Research Corporation
ENVO	Environment Ontology
FAIR	Findability, Accessibility, Interoperability, and Reusability
FAPESP	São Paulo Research Foundation
FIP	FAIR Implementation Profile
FsF	FAIRsFAIR
GBIF	Global Biodiversity Information Facility
GloBi	Global Biotic Interactions
GN	GeoNames
IA	Information Architecture
IAPE	Institute of Applied Economic Research
IEC	International Electrotechnical Commission
IGAD CoP	Improving Global Agricultural Data Community of Practice
INSPIRE	Infrastructure for Spatial Information in Europe
IPEA	Institute of Applied Economic Research
IUCN	International Union for Conservation of Nature and Natural Resources
KG	Knowledge Graph
KALRO	Kenya Agricultural and Livestock Research Organization
KOS	Knowledge Organization System
LLM	Large Language Model
LOD	Linked Open Data
LOV	Linked Open Vocabularies
LPT	Livestock Product Trait Ontology
Mapa	Ministry of Agriculture, Livestock, and Supply
MEAT-T	Meat Thesaurus
MRLs	Maximum Residue Limits
NCM	Nomenclatura Comum do MERCOSUL
NPS	Net Promoter Score

NRCS	Natural Resources Conservation Service
OBIS	Ocean Biogeographic Information System
OFC	FAIR compliance score
OGC	Open Geospatial Consortium
OGP	Open Government Partnership
OFPE	Ontology for Food Processing Experiment
OMV	Ontology Metadata Vocabulary
ODP	Ontology Design Pattern
OTUs	Operational Taxonomic Units
OWL	Web Ontology Language
PPI	Plant-Pollinator Interactions
PLDN	Platform Linked Data Netherlands
RAG	Retrieval-Augmented Generation
RDA	Research Data Alliance
REBIPP	Brazilian Plant-Pollinator Interactions Network
RDF	Resource Description Framework
RO	Relations Ontology
SDO	Schema.org
SHs	Species Hypotheses
SKOS	Simple Knowledge Organization System
SUS	System Usability Scale
TDWG	Biodiversity Information Standards
UDC	Universal Decimal Classification
UFO	Unified Foundational Ontology
UML	Unified Modeling Language
UNSD	United Nations Statistics Division
URI	Uniform Resource Identifier
USDA	United States Department of Agriculture
USP	University of São Paulo
UX	User Experience
XSLT	Extensible Stylesheet Language Transformations

Part I

Setting the Stage

Introduction to Part I

In the age of digital agriculture, achieving seamless data integration across diverse platforms is vital for unlocking the full potential of data-driven insights. Part I of this dissertation lays the theoretical and methodological groundwork for addressing the challenge of semantic interoperability in agricultural contexts.

Chapter 1 sets the stage by defining the scope of the problem: data heterogeneity and semantic ambiguity obstruct effective data exchange and reuse. It underscores the importance of layered interoperability, particularly at the semantic level, and outlines the core issues stemming from inconsistent vocabularies and misaligned data structures.

Building on this foundation, Chapter 2 introduces a structured semantic interoperability framework modeled in BPMN. The framework guides practitioners through the stages of goal definition, data preparation, ontology alignment, and deployment, providing a process-driven approach to harmonizing agricultural datasets.

Chapter 3 complements the previous discussions by presenting a conceptual model for FAIR metadata schemas, developed using OntoUML. This model operationalizes the FAIR principles within metadata design, facilitating semantic precision and reusability. The chapter emphasizes not only conceptual clarity but also practical implementation through annotated examples.

Together, these chapters establish a comprehensive base for developing scalable and reusable data infrastructures in digital agriculture, setting the stage for more advanced implementations and evaluations in subsequent parts of the dissertation.

Part II and III of this dissertation present two realistic case studies that illustrate the development and implementation of the proposed framework for semantic interoperability. The first case study focuses on agricultural price index datasets and is explored in significantly more detail. This reflects its central role in the research: it involved extensive processes such as data harmonization, the design of new domain-specific semantic artifacts, and the application of Large Language Model (LLM) both to support ontology construction and to facilitate querying of a Knowledge Graph (KG). As such, it served as the leading case study driving much of the methodological innovation in this dissertation.

The second case study, in contrast, concerns the domain of plant–pollinator interactions and is described more concisely. While the author has been involved in projects related to this case study since 2020, many foundational elements of semantic interoperability (such as metadata schema design and ontology development) had already been addressed in prior collaborative work, which is referenced and summarized in Chapter 11. Additionally, the author's role in this second case study was more indirect. These factors account for the reduced length and presentation detail, although its inclusion remains relevant to this dissertation to demonstrate the framework's adaptability across domains with varying levels of semantic maturity.

CHAPTER 1

Introduction

Chapter Overview

This chapter introduces the core concepts and motivations behind this dissertation, focusing on the challenges of semantic interoperability in the context of digital transformation in agriculture. It begins by discussing how data incompatibility (stemming from inconsistent formats, ambiguous semantics, and heterogeneous systems) limits the value of digital data. The chapter defines the different layers of interoperability in Section 1.2, with emphasis on semantic interoperability, and outlines common sources of semantic conflict such as naming mismatches and structural discrepancies in Section 1.3. It highlights the role of ontologies, metadata schemas, and schema languages like RDF and OWL in resolving these challenges. Special attention is given by Section 1.5 to semantic interoperability in agriculture, where the need for harmonized data is especially pressing. A case from Brazil illustrates real-world metadata inconsistencies that motivate the framework proposed in this work. The chapter concludes by stating the research hypothesis, objectives, and design science methodology, setting the foundation for the development and evaluation of the semantic interoperability framework detailed in the following chapters.

1.1 Background

Digital transformation is a phenomenon that arose from the development and dissemination of information technologies, which directly impacted the operational processes of public and private organizations, governments, and industries of all kinds [1]. Digital transformation aims to enhance productivity and improve people's lives [1].

In today's digital transformation landscape, the proliferation of information systems has brought about a significant challenge: the problem of data incompatibility. As organizations and industries adapt to the ever-evolving technological landscape, the exponential growth of digital data and the diversity of information systems has led to a complex and often insurmountable hurdle. The seamless exchange and utilization of data across disparate platforms have become imperative for streamlined operations and informed decision-making. However, the lack of standardized formats and harmonious data structures has engendered a landscape where data incompatibility impedes efficient communication and collaboration.

Like other sectors, agriculture has embraced digital transformation to enhance processes and increase production value. Terms such as digital agriculture, e-Agriculture, and Agriculture 4.0 encompass the application of diverse technologies, including big data, mobile technologies, robotics, remote-sensing services, virtual and augmented reality, distributed computing, the Internet of things, and adaptive systems, in agriculture [2, 3, 4, 5]. These efforts are crucial as agriculture remains one of the least digitized sectors worldwide [6], presenting an opportunity for digitization benefits.

Despite the vast potential for utilizing data in agriculture to enhance decision-making and innovation, existing barriers hinder access to certain datasets. These barriers may include unclear use licenses, formats that impede easy retrieval and reuse, and the absence of proper semantic definitions, which can lead to erroneous inferences or misuse of the data [7]. It is crucial to address these challenges to leverage the benefits of data in agriculture fully. This ensures datasets are accessible, appropriately licensed, and paired with meaningful semantic descriptions. The FAIR data principles [8] offer valuable guidance in this regard.

The FAIR data principles address various challenges and provide a comprehensive framework to make data Findable, Accessible, Interoperable, and Reusable [8, 9, 10]. These principles propose guidelines to improve the discoverability and accessibility of data, making it easier for researchers and stakeholders in fields such as in agriculture to locate and access relevant datasets. Recognized as valuable guidelines for data and metadata management, the FAIR principles aim to facilitate data retrieval on the web through persistent identifiers, ensure seamless data access, and utilize metadata for efficient information exchange between systems, all while promoting data reuse [8, 9, 10].

Adopting structured data, enriched with semantic descriptions and standardized formats, enhances data processing efficiency for both machines and humans. For example, LLMs can significantly benefit from such structured data, particularly when presented as Knowledge Graphs (KGs). This structure aids LLMs in gaining domain-specific knowledge, enhancing analysis accuracy, decisiveness, interpretability, and more [11]. For human users, structured data streamlines tasks, as the data meaning can be discerned directly from the dataset and its associated metadata. This eliminates the need to consult external sources for comprehension. Reports from CrowdFlower [12] and IBM [13] indicate that data scientists predominantly allocate their time to "data wrangling" – the collection, cleaning, and organization of data to optimize it for analysis or machine learning applications, also highlighted by [14].

In the intricate domain of Agriculture, adhering to these principles can empower the agricultural community to enhance data transparency, foster collaboration, and make informed decisions, thereby propelling progress in the field. Such adherence paves the way for new job opportunities and bolsters collaborative endeavors among stakeholders, yielding richer insights and well-informed decisions [6]. By championing open data principles, we establish a robust foundation that maximizes the advantages of digitization, positioning the agricultural sector to thrive in an increasingly data-driven world. Many countries and organizations worldwide have started working to promote open data in agriculture, as indicated by [5].

In Brazil, the movement towards open data in the public sector is an integral part of the Open Government Partnership (OGP). The OGP's initial milestone is the Law on Access to Information [15], which ensures open and free access to data generated by public administration institutions [16]. The partnership's inception dates back to 2011 and has been overseen by the Comptroller General of the Union (CGU) since then. Notably, in 2018, the 4th National Open Government Action Plan was introduced, encompassing the creation of an Open-Data Ecosystem. This involved opening data from all levels of public administration in Brazil, including federal, state, and municipal governments [17].

This approach to open data in the public sector reflects Brazil's commitment to transparency and promoting collaboration and innovation through data-driven governance. By embracing open data practices, the country aims to enhance public service delivery, empower citizens with access to relevant information, and foster an environment conducive to social and economic development. The Open-Data Ecosystem is expected to facilitate the sharing and reuse of valuable data across governmental entities, enabling data-driven decision-making, evidence-based policies, and collective progress. This effort is expected to foster change and accountability within the public sector.

Open data should consistently be paired with detailed semantic descriptions, ensuring users can comprehend and scrutinize the dataset via explicit metadata. Recognizing the significance of adhering to optimal metadata practices, especially in light of the FAIR Data principles, the 5th Brazilian National Open Government Action Plan has taken measures to address this [18]. A key objective of this plan was to set forth protocols for data disclosure, including the Brazilian agricultural sector [18].

Aligned with OGP, the GoFAIR Brazil Initiative established the GO FAIR Agro Brazil Network to apply the FAIR principles to Brazilian agricultural data, thereby fostering improved data sharing and reuse [19, 20]. The GO FAIR Agro Brazil Network is expected to unlock the potential of open data to drive innovation, evidence-based decision-making, and sustainable agricultural practices throughout the country. As a result, this initiative has the potential to revolutionize agricultural research, promote collaboration, and empower stakeholders with valuable insights for driving the sector's future.

1.2 Interoperability

Interoperability denotes the seamless ability of diverse systems to communicate, share, and utilize data without conflict [21, 22, 23, 24]. Interoperability is described in the literature as having many layers, such as technical, syntatic, semantic, structural, and organizational interoperability.

Technical or system interoperability: this layer of interoperability primarily concerns the foundational mechanisms that facilitate communication between distinct systems [25]. It encompasses the hardware and software components, ensuring that data can be transmitted seamlessly from one system to another [25]. It encompasses data exchange protocols, data formats, and connectivity, without delving into the actual meaning or semantics of the data [24]. An example could be "moving data from a USB stick to a computer" [25].

Syntactic interoperability: delves into the structural aspects of data exchange. It ensures that the data exchanged between systems adheres to specific formats, allowing for consistent parsing and interpretation [23, 26]. The preservation of data structure is paramount, often achieved through the adoption of standardized data formats that provide a common syntactical framework [26].

Semantic Interoperability: ensures the preservation and consistent interpretation of data meaning across systems. This level addresses potential discrepancies in terminology and data definitions, ensuring that the information precise meaning remains consistent (without ambiguity) and comprehensible across diverse systems or applications [24]. Semantic interoperability is sometimes related to syntactic interoperability, as variations in data format may have implications for the data semantic interpretation [27].

Structural Interoperability: emphasizes the intricacies of data exchange through predefined conceptual models [26]. It ensures that diverse data structures are coherently understood and translated across systems. The focus is on maintaining the integrity of data organizational format (e.g., relationships between classes and attributes) during exchanges [26].

Organizational Interoperability: Refers to the alignment of business processes, legislation, policies, responsibilities, and expectations among organizations to achieve mutual goals [25, 27]. It involves documenting and integrating business processes and relevant information exchanges [27]. Some authors classify certain aspects of this layer as distinct layers in their own right, such as Process Interoperability [25] and Legal Interoperability [28].

1.3 Semantic Interoperability

In the digital age, the transformation of implicit knowledge, which is the innate human understanding of the real world, into explicit knowledge is a pivotal process [29]. This explicit knowledge is articulated through information models that serve as structured representations of data about the real world [29]. The work of [29] delves into this transformation, emphasizing the role of semantics in bridging the gap between human cognition and machine interpretation. It underscores the significance of context in clarifying meaning and the challenges of mapping real-world phenomena into comprehensible, structured data models.

Semantic interoperability emphasizes preserving data meaning and context during data exchanges between systems [26]. This entails that both the data consumer and the data provider share a unified understanding of the data meaning [21]. Heterogeneities in information systems, which can manifest as either structural or semantic discrepancies, pose challenges to semantic interoperability [30].

1.3.1 The Semantic Interoperability Problem

The broken telephone metaphor (Fig. 1.1) aptly illustrates the semantic interoperability problem in data exchange between applications. Just like in the children's game, where a message loses its original meaning and structure as it passes from people to people, data can undergo unintended transformations and lose its semantics as it moves across information systems.

In the context of data science, semantics provide the vital contextual meaning that allows humans to understand and use the information effectively, which is given in terms of metadata. However, if the accompanying metadata is insufficiently descriptive, it can lead to challenges in comprehending and reusing data [8].

As data volumes continue to increase, humans struggle to process vast amounts of information as rapidly as computers can [8]. Consequently, computer agents are increasingly employed to discover, analyze, and integrate data. For these agents to function properly, data must be managed to enable automated usage, which is a priority for data producers, consumers, and repositories [8].

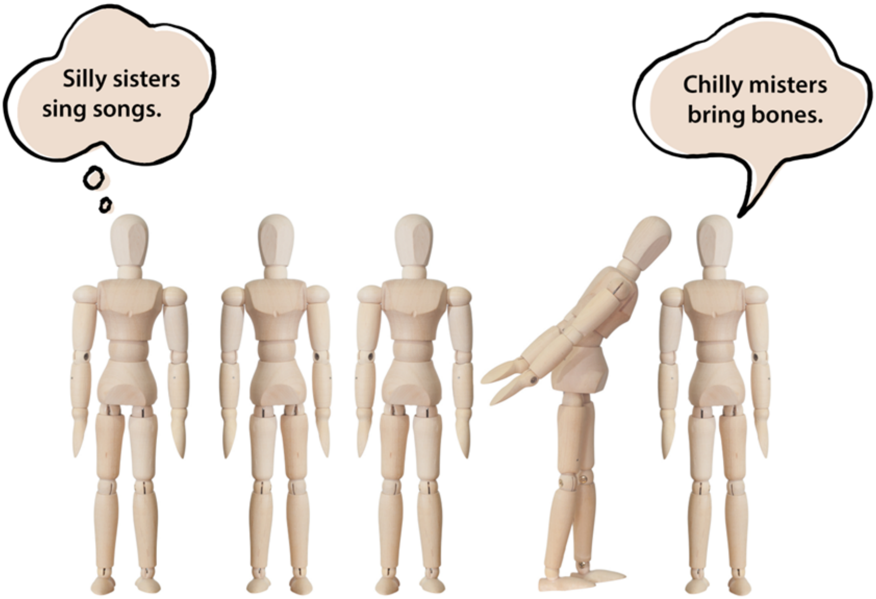


Figure 1.1 The Broken Telephone Game Metaphor.

To achieve interoperability across information systems, it is crucial to store data in well-structured datasets that are accurate, reliable, controllable, and verifiable [31]. The correct interpretation of data also depends on consumers' and producers' shared understanding of the meaning of data fields, which can be achieved through a set of descriptive attributes [31].

Semantic interoperability, the focus of this research, ensures that data exchanged within information systems makes sense to both the data requester and the data provider. It guarantees that the exchanged information is well understood by both parties, preserving the intended meaning and context of the data [21, 22].

Achieving semantic interoperability requires a mutual agreement on the shared data meaning. This entails choosing identical semantic artifacts, such as metadata schemas and ontologies, to describe the data. However, the challenge lies in the multiple layers of the semantic interoperability problem, including conflicts in elements' definitions (such as naming, identification, and constraint conflicts) and conflicts in domain representation (such as abstraction-level incompatibilities, multilateral matches, metamodel-level discrepancy, and domain coverage) [30].

Structural heterogeneities arise at the model level (which could also be considered a problem of structural interoperability), manifesting as incompatibilities in model elements like entities, attributes, and relationships [30]. Additionally, they can be observed in the organization of these elements within a specific domain [30]. Such heterogeneities can be categorized into two primary types: a) conflicts stemming from variations in elements' definitions, which encompass discrepancies in naming, identification, and constraints; and b) conflicts in domain representation, which occur when domain experts structure model elements differently, even when addressing the same domain [30]. These conflicts arise from the inherently subjective nature of knowledge modeling and representation, which is heavily influenced by context, domain, and specific application [32].

Conflict on elements' definitions may be:

- **Naming conflicts**: refer to the scenario where different metadata schemas use different terms to describe the same concept or real-world entity. For instance, one schema might use the term "DOB" while another uses "Date of Birth", both referring to the same concept. [30].
- **Identification conflicts**: arise when there are differences in how elements in metadata schemas are uniquely identified. Some schemas might use simple names as identifiers, while others might use more qualified identifiers that include namespaces or other distinguishing features [30].
- **Constraint conflicts**: occur when there are issues importing definitions from one system schema into another system. This could be due to differences in how constraints are defined or enforced in the two systems [30].

Conflicts of domain representation may be:

- **Abstraction-level incompatibilities**: mismatches that arise due to differing levels of abstraction in the represented knowledge or data [30]. It could happen between two data models, where one describes things in a very detailed, granular manner, while the other describes things at a high, generalized level [30].
- **Multilateral matches**: when an element of a schema has multiple correspondent elements in another schema, and vice versa [30].
- **Metamodel-level discrepancy**: when an element of a schema has no direct equivalent in another schema [30].
- **Domain coverage**: when two schemas are built on the same domain of knowledge but do not have common concepts to represent the same domain [30].

As for semantic heterogeneities, we understand the semantic differences among models. The model semantics depends on the domain and the interpretations of domain entities to model elements. The semantic domain gives meanings to each model's element through linguistic expressions [30]. Semantic heterogeneities may be:

- **Domain conflict**: when domains overlap, aggregate others, or are incompatible [30].
- **Terminological incompatibilities**: when the same domain concept is mapped to model elements with different names (synonym conflict) or when distinct domain concepts are mapped to model elements with the same names, i.e., homonym conflicts [30].
- **Data encoding conflict**: incompatibilities in the format of the values assigned to the metadata fields, e.g., date=01.01.2007 or date=2007/01/01, or conflicting units of measure are used to fill in metadata values [30].
- **Controlled vocabulary conflict**: the same as the terminological conflict, but happens when conflicting controlled vocabularies are used to fill in metadata fields. These controlled vocabularies could be ontologies with different namespaces for the same domain, for instance [30].

Semantic interoperability conflicts encompass a myriad of complexities, interacting with various artifacts. Amongst these artifacts, upper ontologies function as the meta-meta-models for data structures. While many conflicts can arise at this juncture, it is also the pivotal point where domain representation disputes ought to be resolved.

Foundational ontologies like the Unified Foundational Ontology (UFO) and the modeling language OntoUML can be instrumental in understanding, structuring, and representing complex and messy data structures in many ways:

- **Conceptual clarification**: Foundational ontologies help distinguish between different types of entities and their relationships. For instance, distinguishing between substantial entities (things that exist in the world) and relational entities (relationships between things) can help in understanding the nature of the data [33].
- **Structural organization**: OntoUML provides a rich set of modeling constructs grounded in the UFO ontology. This allows for a more precise representation of the domain, capturing nuances that might be overlooked in traditional UML [34].
- **Semantic rigor**: Foundational ontologies ensure that the concepts and relationships in the model are semantically sound and consistent. This can be especially valuable when dealing with data that might seem ambiguous or contradictory [35].
- **Domain exploration**: By modeling the domain using OntoUML and UFO, one can explore different facets of the data, uncover hidden relationships, and identify potential inconsistencies or gaps [33, 36].
- **Facilitating reuse and interoperability**: The use of OntoUML and UFO supports the reuse of existing core ontologies, many of which

have already been developed based on these foundations. This enables modelers to build upon consistent and coherent ontologies, promoting interoperability and reducing redundancy. Reuse is intrinsic to ontology engineering and represents a significant advantage when adopting these frameworks [37].

All these characteristics allow to build of a well-structured ontology that may serve as a basis for data integration, ensuring that data from different sources or domains can be coherently merged and understood in a unified context. The lack of foundational ontologies can lead to data structures that are challenging to comprehend and reuse.

Choosing the right schema definition language is crucial for determining how data is represented and how effectively it can be shared or integrated with other systems – essentially, its interoperability. Different languages have unique features tailored to specific requirements. For example, Resource Description Framework (RDF) serves as a "standard model for data interchange on the Semantic Web" [38]. In contrast, Web Ontology Language (OWL) was created to depict "rich and complex knowledge about things, groups of things, and relations between things" [39]. Meanwhile, Simple Knowledge Organization System (SKOS) is ideal for representing thesauri, classification schemes, and taxonomies [40]. While each language boasts its strengths, they also have inherent limitations. However, despite these constraints, they are notably more expressive than plain HTML in preserving data semantics.

In a more granular layer, metadata schemas, such as the Data Catalog Vocabulary (DCAT) and Dublin Core, encompass classes and properties that can be instantiated with data representing real-world entities. These schemas play a pivotal role in resolving naming conflicts at the implementation level by standardizing column names in databases, data types, and other attributes, for instance.

The choice of which schema to use must be made judiciously for each case. In the case study 1 described in this dissertation, for instance, we developed the Agriculture and Livestock Metadata Elements Set (Almes Core) [41] to fulfill a gap in the domain of agricultural trading knowledge representation. While it draws inspiration from established metadata schemas like DCAT, Dublin Core, and Schema.org, it introduces unique terms designed for the domain.

In summary, Fig. 1.2 illustrates how semantic technologies can be integrated to effectively address the challenge of semantic interoperability across multiple layers of information exchange [42]. Starting from the topmost layer, ontologies and metadata schemas offer a unified vocabulary that applications should adopt to standardize data descriptions. Embracing an ontology ensures that all stakeholders involved in data creation concur on a universally accepted meaning for data within a specific domain. Conceptual models, on the other hand, shed light on an application's data structure. They capture the relationships among different system components (for instance, the connections between tables in an SQL database). Such relationships can be articulated using modeling languages like Unified Modeling Language (UML), OntoUML, and Enhanced Entity–Relationship (EER) model. At the implementation tier, systems responsible for data storage should consistently process ontology and metadata terms (e.g., properly resolve URIs) and adhere to the conceptual model.

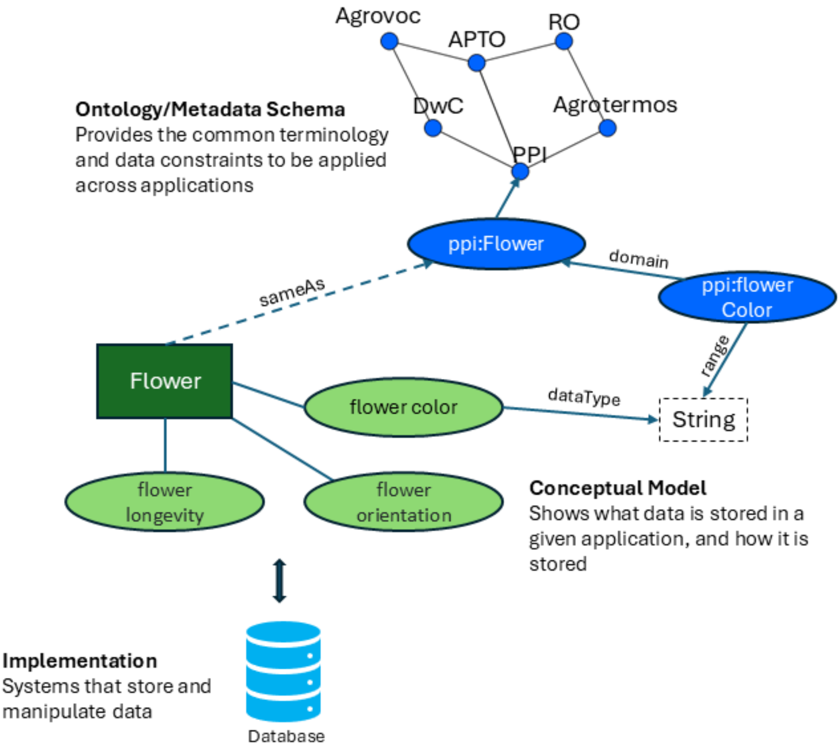


Figure 1.2 Information System Integration Using Ontologies (Adapted from Keet [42]).

1.4 Related Work on Semantic Interoperability

While semantic interoperability has been widely studied, the development of domain-specific frameworks remains fragmented across sectors. Notable advancements have emerged in domains such as healthcare, life sciences, and geospatial data. These sectors offer valuable insights and methodologies applicable to agriculture.

In healthcare, the Fast Healthcare Interoperability Resources (FHIR) standard developed by Health Level 7 (HL7) [43] provides structured, semantically-rich APIs to exchange patient data. The SNOMED Clinical Terms (SNOMED CT) have supported semantic annotation and retrieval in Electronic Health Records (EHRs) [24]. FHIR defines the structure for clinical data exchange, while SNOMED CT provides the semantic layer through standardized clinical terminologies. Used together, they form the backbone of semantic interoperability in modern healthcare systems [44, 45, 46].

The biodiversity informatics community has made substantial progress in semantic interoperability, offering a mature and coordinated set of standards and practices that can inform similar efforts in agriculture. Central to this ecosystem is the Darwin Core (DwC) standard, developed and maintained by Biodiversity Information Standards (TDWG), which provides a structured vocabulary for describing species occurrence, taxonomic classification, and associated ecological metadata [47]. Widely implemented by global infrastructures such as Global Biodiversity Information Facility (GBIF) [48], DwC ensures consistent syntactic and semantic representation of biodiversity data. To complement DwC, the Ecological Metadata Language (EML) supports detailed dataset-level documentation, capturing aspects such as data provenance, collection methods, and temporal-spatial coverage [49]. These two standards, along with packaging mechanisms like the Darwin Core Archive (DwC-A) [50], enable efficient integration and FAIR-compliant publication of large-scale biodiversity datasets. Beyond structural standards, the biodiversity domain has also begun incorporating semantic web technologies, including RDF and OWL-based ontologies. For instance, Darwin-SW (DSW) has been developed as a lightweight RDF model to complement DwC [51], while the Environment Ontology (ENVO) [52] and Biological Collections Ontology (BCO) [53] enrich ecological and sample-level metadata with formal semantics. These advances facilitate cross-repository data reuse and machine reasoning, enhancing discovery and integration capabilities. Furthermore, initiatives like DataONE [54] and Ocean Biogeographic Information System (OBIS) [55] exemplify how standardized vocabularies and controlled metadata practices can support global-scale, semantically interoperable data ecosystems.

Semantic interoperability has been a central concern also in the geospatial domain, especially with the increasing demand for integrating heterogeneous spatial datasets. A foundational effort in this domain is the Open Geospatial Consortium (OGC)'s development of GeoSPARQL, a W3C-aligned standard that extends SPARQL for querying geospatial data using topological relations (e.g., within, intersects) expressed through standard geometry representations [56, 57]. GeoSPARQL has been widely adopted in both research and government platforms to facilitate semantic querying of linked spatial datasets. Studies such as Jovanovik et al. [58] have benchmarked RDF triple stores for GeoSPARQL compliance, showing significant variations in performance and standards support across platforms. Complementing these efforts, tools like GeomRDF convert traditional GIS formats into RDF with fine-grained spatial semantics, enabling publishing and integration of geospatial linked data while maintaining geometry structure [59]. National and regional agencies have begun adopting such technologies to align with broader Infrastructure for Spatial Information in Europe (INSPIRE) directives [60] and FAIR data principles [8]. Other recent innovations include federated GeoSPARQL querying frameworks that optimize source selection using spatial metadata [61], and semantic reasoning frameworks for geospatial policy support using OWL and PROV-O [62]. The evolution of geospatial semantics is heading towards generating queries from natural language questions using LLMs [63, 64].

Despite substantial progress across domains, significant challenges in achieving semantic interoperability remain. Persistent issues include ontology alignment, terminological inconsistencies, model-level conflicts, and the emerging complexity of generating effective KG queries using LLMs. This dissertation draws upon many of the techniques and frameworks proposed in related domains, integrating them throughout its chapters. Furthermore, it builds upon and extends the state-of-the-art semantic artifacts already developed in the agricultural sector.

1.5 Semantic Interoperability in Agriculture

As emphasized by Khatoon and Ahmed [65], achieving semantic interoperability offers substantial benefits to diverse stakeholders across the agricultural value chain, including farmers, traders, decision makers, and many other actors. Raw data collected from soil sensors, drones, and local weather stations through precision agriculture may offer limited advantages as they lack meaningful organization [65]. However, when this data is contextualized, given meaning, and integrated with additional data sources, it becomes valuable information that can drive better decision-making in agriculture [65].

Publishing agriculture data with detailed semantic descriptions holds the potential to empower stakeholders in making more informed and appropriate actions to enhance food production [3]. FAO and ITU [5] emphasize that data interoperability is a powerful tool to unlock the value of agricultural data generated by various providers. Several countries, including China, Australia, and Afghanistan, are already embracing data interoperability as part of their e-government initiatives, recognizing its significance in agricultural advancement [5]. Despite these advancements, achieving global semantic interoperability in digital agriculture remains a challenge that requires continuous efforts and collaboration.

The domain of agriculture trading in Brazil, chosen as the main case study (Case Study 1) for this dissertation, faces significant semantic interoperability challenges, such as the variability in metadata terminology among data providers, conflicting data models, etc. For example, three datasets from key data providers in Brazil's agriculture commodities sector, namely the Center for Advanced Studies on Applied Economics (CEPEA)[1], the Institute of Applied Economic Research (IPEA[2]), and The National Supply Company (CONAB)[3]), exhibit distinct data models and metadata terminology to describe similar data. For instance, when describing the temporal occurrence of data, CEPEA uses the term 'periodicidade' (periodicity) in Portuguese, whereas IPEA uses 'frequência' (frequency) for the same concept. This inconsistency in terminology creates a naming conflict between CEPEA and IPEA, as we see in Fig. 1.3.

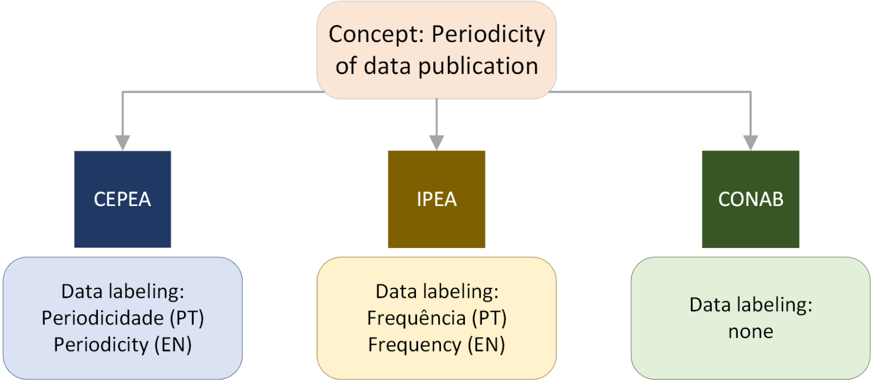

Concept: Periodicity of data publication
CEPEA
IPEA
CONAB
Data labeling: Periodicidade (PT) Periodicity (EN)
Data labeling: Frequência (PT) Frequency (EN)
Data labeling: none

Figure 1.3 Example of Semantic Incompatibilities within Datasets.

CONAB lacks a specific metadata descriptor for this concept altogether and only offers options to filter data based on monthly or weekly publications. This disparity results in a meta-model level incompatibility between CONAB and the other data models, as it lacks the necessary semantic artifacts to effectively describe the temporal occurrence of data.

Furthermore, a critical concern is the lack of explanation regarding the meaning of metadata terms in both the datasets and the data providers' websites. Absence of clear documentation makes it challenging to reuse the data effectively, as users may encounter difficulties in ascertaining the precise significance of the labels used in the datasets. Without comprehensive explanations, there is a significant risk of misinterpretation or incorrect utilization of the data. Addressing this issue by

---

[1] `https://www.cepea.esalq.usp.br/br`

[2] `http://www.ipeadata.gov.br/Default.aspx`

[3] `https://sisdep.conab.gov.br/precosiagroweb/`

providing comprehensive and standardized explanations for metadata terms is essential to enhance data usability and facilitate seamless data interoperability within the domain.

1.6 Hypothesis

Implementing a framework for semantic interoperability that integrates metadata schemas, conceptual models, and ontologies will lead to enhanced semantic interoperability between agriculture datasets. This integration is expected to resolve conflicts related to concept definitions and domain representations, thereby facilitating data exchange, integration or federated search, and understanding within the agricultural domain.

Metadata schemas are expected to play a crucial role in resolving conflicts related to elements' definitions, ensuring standardized and clear representations across datasets [66]. Conversely, the utilization ontologies addresses conflicts arising from domain conceptualization and representation [67], establishing a coherent and consistent knowledge organization specific to the agricultural sector.

The FAIR data principles are also expected to be a reference for best practices when developing and applying this framework [8]. Metadata schemas provide terms with semantic and syntactic definitions for describing data. When this schema is findable and accessible on the Web, anyone can reuse it to describe their datasets. When two or more distinct datasets use the same metadata elements, they can be integrated. The semantic integration of heterogeneous datasets is critical to the proper functioning of many Web applications that currently exist and that need to query these datasets [68, 69].

1.7 Objective

This dissertation aims to advance semantic interoperability in the agricultural domain by designing a practical and adaptable framework aligned with the FAIR data principles. The research integrates conceptual modeling, metadata schema design, ontology engineering, and tooling to support semantic data management across diverse agricultural datasets.

1.7.1 Main objective

To propose and implement a comprehensive semantic interoperability framework (comprising conceptual models, metadata schemas, domain ontologies, supporting tools, and the necessary implementation processes) that enables the integration, semantic annotation, and reuse of heterogeneous agricultural data in accordance with the FAIR principles.

1.7.2 Specific Objectives

To achieve the main objective, the following specific objectives were pursued:

- Develop a conceptual model and a set of design guidelines for creating FAIR-compliant metadata schemas.
- Design and publish domain-specific semantic artifacts (e.g., metadata schemas and ontologies) that address key gaps in existing agricultural data representations.
- Create a suite of reusable tools and resources to support semantic interoperability workflows, including web applications, Python notebooks, and LLM-based system prompts for specific annotation and integration tasks.
- Apply and test the framework in two real-world case studies, each focused on a distinct subdomain within agriculture.
- Assess the framework's effectiveness using FAIR self-assessment instruments, ontology validation methodologies, and AI-assisted data retrieval evaluations.

1.8 Research Design

We built the framework for semantic interoperability as an artifact, based on Design Science principles. In Design Science, an artifact refers to a novel, purposeful, and innovative solution that addresses a specific problem or fulfills a particular need. It can take various forms, such as algorithms, models, frameworks, prototypes, or complete systems [70, 71].

In Design Science, the prevailing premise is that the construction and evaluation of purposeful artifacts can facilitate the generation of knowledge, particularly when these artifacts are designed to address distinct challenges within practical domains [71]. It is imperative that these artifacts are accompanied by comprehensive documentation detailing their utilization, operational guidelines, and provenance [70]. These attributes are characteristics inherent to Design Science Research. According to Wieringa [71], the focal point of Design Science pertains to the systematic study of artifacts within their contextual environments. The primary research endeavors in this domain involve the meticulous examination and refinement of these artifacts within their respective contexts [71]. Fig. 1.4 shows the components of an artifact.

The application of Design Science in this research aimed at answering the following design question: **How can a semantic interoperability framework be designed to support the FAIR principles and resolve semantic conflicts in heterogeneous agricultural datasets?** To answer this question, each component of the research artifact was mapped to

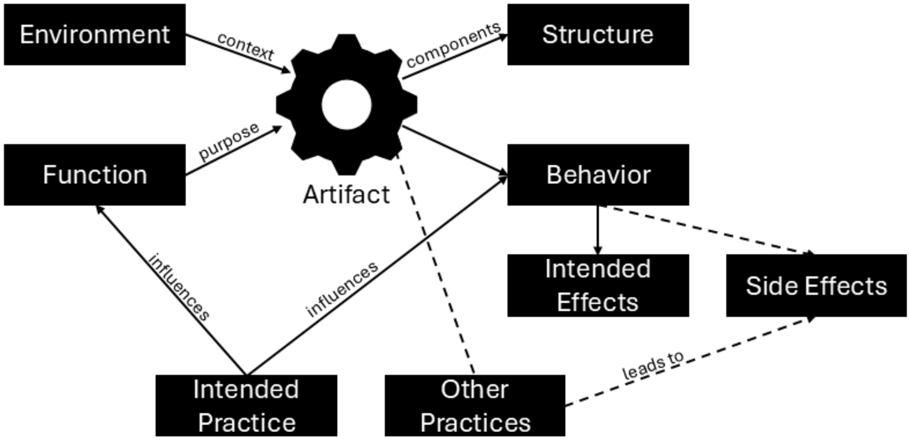


Figure 1.4 The Elements of an Artifact (Adapted from Johannesson and Perjons [70].)

the elements proposed by Johannesson and Perjons' Design Science framework (see Figure 1.4) as follows:

- **Environment**: Brazilian agriculture trading market.
- **Function**: preserve the semantics of data across different information systems produced by organizations related to trading in agriculture.
- **Structure**: the components of the artifact include metadata schemas, ontologies, algorithms for data annotation and modeling, and a Knowledge Graph.
- **Behavior**: the artifact can describe the meaning of a concept in a given dataset and also show the equivalent term for that same concept in another data set. It also allows systems to exchange information and preserve the semantics of this information.
- **Intended effects**: establishing a shared understanding of data meaning across diverse datasets; ensuring that humans can interpret the data without additional research, as the provided semantic descriptions are sufficient; enabling machines to process and reason over the data through its representation in a formal knowledge structure.
- **Side effects**: misconception of the data definitions may lead to incorrect data interpretation and use.
- **Intended practice**: using the semantic interoperability framework to solve compatibility issues in Brazilian agriculture trading datasets.
- **Other practice**: modification of the data publication process of the data providers that adopt the new metadata schema.

It is important to note that this research followed a hybrid approach to framework development, combining both bottom-up and top-down strategies. The bottom-up component was essential for identifying scope, requirements, and practical constraints. This emerged from direct engagement with real-world datasets, domain experts, and institutional partners in this dissertation case studies. Meanwhile, the top-down component grounded the framework in well-established theories, principles, such as FAIR, and modeling practices, e.g., OntoUML and Business Process Model and Notation (BPMN). Together, these strategies ensured both contextual relevance and conceptual rigor in the final framework design.

according to Johannesson and Perjons [70], Design Science research typically follows one of two approaches: (A) developing a general-purpose artifact intended to solve problems across a broad range of contexts, or (B) creating an artifact to address a specific problem in a well-defined setting, with the intention of generalizing the insights and lessons learned to broader domains. In this dissertation, we adopted the second approach. Initially, the semantic interoperability framework was designed by drawing upon prior experience from related domains and projects, as well as insights from the relevant literature. The framework was then applied to Case Study 1, focused on agricultural commodity data in Brazil. Throughout this application, several iterations were required to adapt the framework to the real-world challenges encountered, demonstrating the iterative nature of the Design Science research methodology. Subsequently, the refined framework was applied to Case Study 2 in a different subdomain of agriculture (agricultural biodiversity) to evaluate its generalizability and capacity to address semantic interoperability issues in other contexts. This two-phase process exemplifies the Design Science second approach [70] of starting with a context-specific solution and using the resulting knowledge to develop more broadly applicable frameworks.

The artifact of this research is highly dependent on human rules, norms, and attitudes. As it has both technical and social components, the artifact of this research might be considered a sociotechnical system [70].

BPMN has been selected as the visual language for representing the framework. We are confident that semantic interoperability is a pervasive phenomenon encountered across various processes. Hence, the adoption of a notation like BPMN for process representation seems appropriate. OntoUML has been employed to provide a formal representation of some framework components, such as the conceptual model for FAIR metadata schemas (Chapter 3).

Each component of the proposed framework was developed through an iterative process of design, evaluation, and refinement. The resulting artifacts, namely metadata schemas, ontologies, conceptual models, and tools, were made openly available through platforms like GitHub, Zenodo, and other open-access repositories, ensuring transparency and reusability. We applied the FAIR data principles not only to evaluatie the FAIRness of the semantic artifacts used in the case studies but also to assess the FAIRness of the data products generated through the framework.

Thus, FAIR principles plyed a central role, guiding both the construction of the framework's artifacts and the end-to-end data processing pipelines applied in the use cases. While we acknowledge ongoing efforts to extend or refine the FAIR principles for semantic interoperability, such as those proposed by [72], this research adheres closely to the canonical formulation by Wilkinson et al. [8], as well as the official guidance provided on the GO FAIR Foundation website [9]. These sources were selected for their foundational role and broad community acceptance. Complementary approaches to evaluating FAIRness, tailored to specific artifact types, were also applied, as discussed in the relevant chapters.

Although Design Science provided the overarching methodological structure for this research, the development of individual components was supported by specialized methods tailored to the unique demands of each task. For instance, the SABiO methodology [73] was adopted to guide the ontology engineering process in Case Study 1. While alternative extensions for SABiO exists, such as SABiOx [74], SABiO was chosen due to its wider adoption and maturity, proven track record in similar projects, and strong familiarity within the research group where this dissertation was developed. Its well-documented process and community recognition further supported its selection. In addition to the SABiO methodology, techniques such as ontological unpacking and tree testing were employed to address specific gaps not fully covered by SABiO, particularly in improving conceptual clarity and enhancing usability from the perspective of diverse stakeholders. These complementary methods contributed to a more comprehensive and user-centered development process. The application and adaptation of these methodological approaches are further detailed in the respective parts of this dissertation dedicated to each component of the framework, respectively.

1.9 Dissertation Structure

Part I of this dissertation continues with Chapter 2, which introduces the Framework for Semantic Interoperability, while Chapter 3 presents a Conceptual Model for FAIR Metadata Schemas.

Part II of this dissertation begins with Chapter 4, which describes the activities of Goal Definition and Data Preparation in Case Study 1. Chapter 5 focuses on the Semantic Annotation activity. Chapter 6 introduces the Agricultural Product Types Ontology, and Chapters 7 and 8 explore innovations in this ontology development, specifically Ontology Design Supported by a LLM and Ontology Testing and Validation Through Usability Metrics, respectively.

In Chapter 9, we present the Center for Artificial Intelligence (C4AI) Knowledge Graph on Agricultural Prices Knowledge Graph on Agricultural Prices (KGAP), followed by Chapter 10, in which we discuss the validation of the case study through LLM-generated queries and a FAIR assessment.

Part III begins with Chapter 11, which examines the application of the framework for semantic interoperability in the WorldFAIR project case study and presents a comparison between Case Study 1 and 2. Finally, Chapter 12 provides reflections, future research directions, and concluding remarks.

CHAPTER 2

Framework for Semantic Interoperability

Chapter Overview

This chapter introduces a structured framework for achieving semantic interoperability in data integration projects. Grounded in both literature and practical case studies, the framework is modeled using BPMN and comprises four key phases:

- Goal Definition (Section 2.3): Identifying data providers, defining project scope, and establishing a shared data template.
- Data Preparation (Section 2.4): Converting datasets into a common tabular structure, standardizing formats, and ensuring completeness.
- Data Semantic Annotation (Section 2.5): Selecting or designing metadata schemas and ontologies, mapping data elements, and applying controlled vocabularies to ensure consistent meaning.
- Validation (Section 2.6): Integrating annotated datasets into shared platforms and running queries to verify semantic consistency. This stage concludes with a FAIR assessment of the processed data.

The chapter presents this framework as a reusable methodology for enhancing semantic interoperability in complex, multi-source data environments, with applications in agriculture and agrobiodiversity explored in later case studies.

2.1 Introduction

Semantic interoperability refers to the ability of different systems to exchange information while preserving the meaning, context, and integrity of the data being shared. It goes beyond simply transmitting data between systems; it ensures that the shared information is interpreted consistently across diverse platforms, maintaining its intended meaning. Achieving semantic interoperability is a comprehensive process involving the integration of various artifacts across numerous activities. These artifacts work together to ensure that the data retains its context and meaning as it moves between different systems and environments.

To represent this concept as a process, we have chosen BPMN as the modeling language. BPMN, widely used for visualizing business processes, provides a structured and intuitive way to map out the flow of activities, decisions, and interactions within a process [75]. In the context of semantic interoperability, BPMN can be applied to depict the series of activities and interactions needed to maintain semantic consistency when exchanging data across diverse platforms. This visual representation allows for a clearer understanding of the steps involved in achieving semantic interoperability and highlights the critical touchpoints where meaning preservation is essential.

The framework consists of four main components, each represented by pools and lanes as per the BPMN standard. These components are described in the following sections: 2.3 Goal definition, 2.4 Data Cleaning, 2.5 Data Semantic Annotation, and 2.6 Data Integration and Semantic Interoperability Validation.

2.2 Framework Design

The concepts that form the foundation of this framework are drawn primarily from the literature, combined with practical experience gathered from various collaborative groups and use cases. These include the following key initiatives:

- Workshops organized by the GO FAIR Agro Brasil Network[1] [19, 20, 76]: These workshops played a pivotal role in shaping the framework by facilitating knowledge exchange and promoting best practices in data interoperability within the agricultural domain.
- Meetings during Brazil's 5th National Action Plan on Open Government [18]: These discussions highlighted the importance of transparency and open data, influencing the integration of open government principles in the framework.
- Dedicated sessions for developing the Almes Core Metadata Schema [41]: The development of this metadata schema informed key aspects of the framework, particularly regarding metadata standards and practices for agricultural data.
- Collaboration with the TDWG Biological Interaction Data Interest Group[2] during the creation of the Plant-Pollinator Interactions (PPI) Vocabulary [77] (from 2020 to 2022): This collaboration enhanced the framework's capability to represent complex biological interactions, crucial for biodiversity data interoperability.

---

[1] `https://go-fair-agro.github.io/`

[2] `https://github.com/tdwg/interaction`

- The WorldFAIR Case Study on Agricultural Biodiversity (from 2022 to 2024) [78, 79, 80, 81]: This case study provided practical insights into managing and sharing agricultural biodiversity data, which were directly applied to the framework.
- Active participation in the Cross Domain Integration (CDI) Working Group [82]: Involvement in this working group significantly influenced the framework by incorporating cross-domain integration practices, ensuring the framework's flexibility and applicability across different domains. Beyond the regular discussions online, we have participated in two Dagstuhl Workshops to discuss the theme: *Research Meeting 23393: Data Documentation Initiative (DDI)-CDI Workshop: Realising Interoperable Data Services in the Metadata Ecosystem*[3], and *Research Meeting 24413: Aligning Technology Architectures with Cross-Domain Metadata Models*[4].

Thus, the development of this framework represents a synthesis of theoretical insights from the literature and empirical research gained through diverse use cases. It reflects both the strategic guidance from academic literature and the practical experience of engaging with real-world data interoperability challenges in agriculture, biodiversity, and cross-domain settings.

2.3 Goal Definition

Data-driven projects requiring semantic interoperability requires the integration of at least two distinct datasets. These datasets are often produced by different individuals, departments within the same organization, or even different organizations. Semantic interoperability is necessary to flow data from various sources to a single platform. To achieve this, several activities should be performed to define the project's goal, as illustrated in Fig. 2.1.

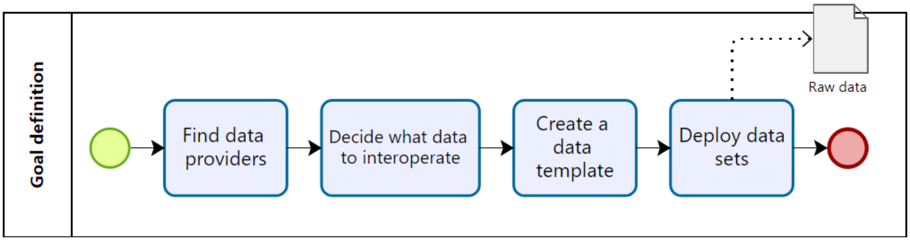


Figure 2.1 Goal Definition Process (Based on [83, 84]).

---

[3] `https://www.dagstuhl.de/seminars/seminar-calendar/seminar-details/23393`

[4] `https://www.dagstuhl.de/seminars/seminar-calendar/seminar-details/24413`

The first task is to `find data providers`, which involves determining the entities responsible for producing the data that we aim to access and reuse. Once the data providers have been identified, the next step is to decide which data to interoperate. This decision should take into account each organization's internal data-sharing policies to ensure that sensitive data is not inadvertently shared. The scope of the data is critical to defining the project's goals and can be guided by questions such as: "What data do we WANT to share?", "What data CAN we share?", and "What are the questions data should be able to answer?". Based on these questions, a `data template` should be elaborated to represent the scope of the expected use for the datasets.

Afterward, each data provider selects the datasets they wish to interoperate and integrate into the platform. The final step is to `deploy the datasets` and make them accessible to the entire project team. For example, datasets can be uploaded to a private cloud, enabling the team to continue with the next phases of the semantic interoperability framework.

2.4 Data Preparation

The first activity in this part of the framework is to `convert datasets to a tabular format`, as shown in Fig. 2.2. If some datasets are in non-tabular formats, such as PDFs or images, they must be converted into a tabular format to enable data manipulations and align with the template defined during the project goal definition. These manipulations include `removing unnecessary columns` that fall outside the scope and goals of the interoperability project, as well as the common data template established in earlier stages.

In some cases, the minimum required information specified in the data template may be missing from the deployed datasets. If this occurs, additional interactions with data providers may be required to `retrieve the missing data` and ensure that the datasets meet the necessary data for interoperability.

Next, `convert data types` (e.g., date fields) from natural language to standardized encoding formats, such as ISO 8601. Afterward, `standardize units of measure` within the datasets to ensure consistency (e.g., converting all time measurements to hours if both minutes and hours are present). Finally, export the harmonized datasets in a tabular format and share them with the entire project team, allowing the project to move forward to the next phase of the semantic interoperability framework.

This process is based on well-established practices in data science pipelines described in the literature [85, 86, 87], which emphasize structured workflows for data preparation, cleaning, transformation, and

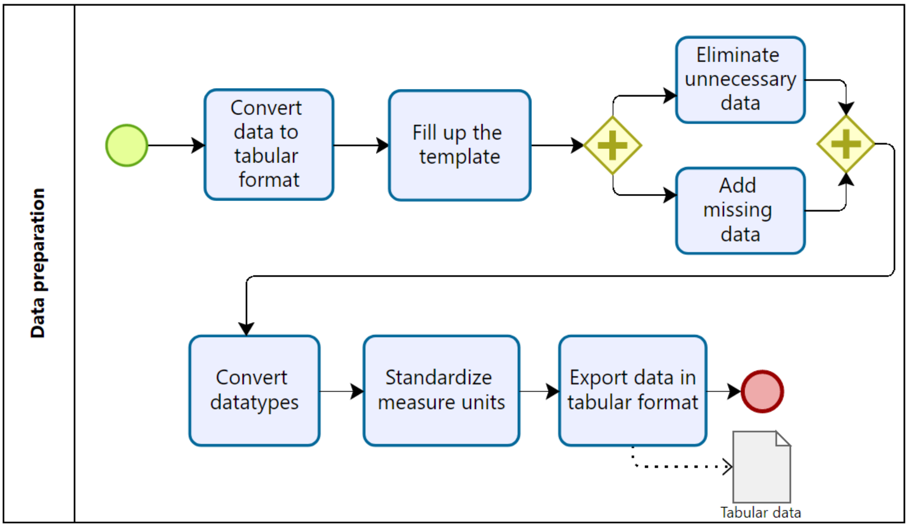

Data preparation
Convert data to tabular format
Fill up the template
Eliminate unnecessary data
Add missing data
Convert datatypes
Standardize measure units
Export data in tabular format
Tabular data

Figure 2.2 Data Preparation Process.

formatting as critical steps for data harmonization.

2.5 Data Semantic Annotation

Semantic annotation is the process of linking parts of an electronic resource (such as text, images) to a formal semantic description, often provided by an ontology. This allows the meaning of the annotated data to be precisely defined, making it easier for machines to understand and process the information. The process typically involves three key components: an ontology, a semantic annotation structure model (which organizes the mappings between the resource and the ontology), and an application (which enables the use of the annotated data for tasks like integration, reuse, or sharing) [88].

Data semantic annotation is the most complex part of this framework. The activities involved are detailed in Fig. 2.3. The process begins with understanding the meaning of each column name in the datasets and listing them. Different datasets often use different column names, but this issue is resolved by mapping them to a metadata schema. The next activity is to `look for a domain metadata schema` that can harmonize the column names. This search may be done in semantic web schema catalogues such as Linked Open Vocabularies[5], OntoBee[6], the Digital Curation Centre's List of Metadata Standards[7] or in domain

---

[5] `https://lov.linkeddata.es/dataset/lov/`
[6] `https://ontobee.org/`
[7] `https://www.dcc.ac.uk/guidance/standards/metadata/list`

specific catalogues, such as the AgroPortal[8]. The selected schema should accurately represent the domain semantics, ensuring that the original meaning of the data is preserved after mapping. If no suitable metadata schema is available, a new one should be created.

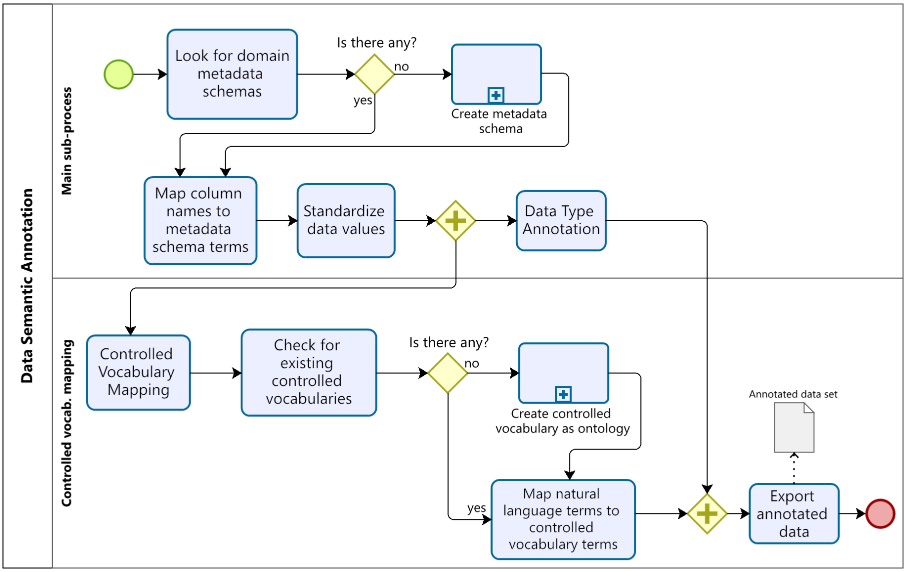

Data Semantic Annotation
Main sub-process
Look for domain metadata schemas
Is there any?
no
yes
Create metadata schema
Map column names to metadata schema terms
Standardize data values
Data Type Annotation
Controlled vocab. mapping
Controlled Vocabulary Mapping
Check for existing controlled vocabularies
Is there any?
no
yes
Create controlled vocabulary as ontology
Map natural language terms to controlled vocabulary terms
Annotated data set
Export annotated data

Figure 2.3 Data Semantic Annotation Process (Based on [88, 89, 90]).

2.5.1 Metadata Schema Specification

To create a new metadata schema from scratch, the FAIR Data Principles [8] can guide the inclusion of elements that ensure the schema is findable on the web, accessible, interoperable, and reusable by others. While the FAIR Data Principles are not mandatory for creating a metadata schema, their application can significantly enhance the schema's reusability. As suggested by [91], metadata schemas, as well as other semantic artifacts like ontologies, can greatly benefit from adhering to these principles.

On chapter 3, we present a more detailed conceptual model explaining how to specify metadata schemas according to the FAIR Data Principles. Within this semantic interoperability framework, developing a new metadata schema should involve at least the activities illustrated in Fig. 2.4.

The process begins with schema `purpose identification`, which involves defining the scope of the schema and determining which part of the domain it should describe. Next, select a core set of elements.

---

[8] `https://agroportal.lirmm.fr/`

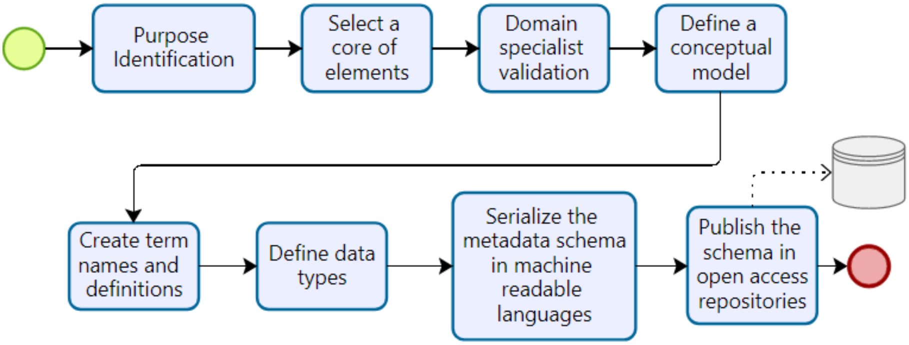

Purpose Identification
Select a core of elements
Domain specialist validation
Define a conceptual model
Create term names and definitions
Define data types
Serialize the metadata schema in machine readable languages
Publish the schema in open access repositories

Figure 2.4 Creating Metadata Schema Sub-Process.

These elements could be found in literature or identified as the names of datasets currently published in the specific domain. Consulting with specialists can also help understand which core elements should be represented in the metadata schema.

Once a list of potential metadata elements is selected, validate these choices with specialists. The next activity is to `define a conceptual model` for the metadata schema. Languages, such as UML and OntoUML, can be used for this task. The conceptual model supports a comprehensive understanding of the metadata schema, including its classes, properties, and relationships.

Finally, provide a detailed description for each metadata element in the schema, including definitions and names for each element. Additionally, specify the expected data types for each element. Once the glossary of terms is generated, serialize the schema relationships using a machine-readable language, such as OWL, RDF, or SKOS. The schema should be published in open access repositories with extensive documentation on how to use it, along with different serializations.

2.5.2 Mapping

After identifying or developing a suitable metadata schema, the next activity is to map the column names to the elements in the metadata schema. After that, the original column names should be replaced with the corresponding metadata element names. This replacing can only be done when there is an exact match between the column ‘concept’ and the metadata element.

Once the columns are renamed, the next step is to consult the metadata schema documentation to understand the required data format. This includes applying primitive data types (e.g., string, literal, float) to all metadata fields in the data set, except for those requiring controlled vocabularies. Controlled vocabulary metadata fields should use terms from an existing controlled vocabulary, such as an ontology, thesaurus, or glossary. If a suitable controlled vocabulary exists for the domain, it should be reused. A new controlled vocabulary should be created if none is available or existing vocabularies do not cover all necessary terms.

In the case studies presented in this dissertation, we modeled controlled vocabularies as OWL ontologies, though they could also be simpler semantic artifacts like glossaries or thesauri. Regardless of the type, each element in the controlled vocabulary should be assigned a unique URI to allow for direct referencing.

2.5.3 Ontology Development

When creating a new ontology for use in controlled vocabulary fields in a dataset, the ontology should be modeled as a type ontology (a classification scheme). This type of ontology is not intended for direct instantiation but rather serves as a classification framework. In the semantic interoperability framework applied in this dissertation, we used the SABIO methodology [73] to develop the ontology for one of the case studies. Fig. 2.5 illustrates the activities for ontology development based on [73].

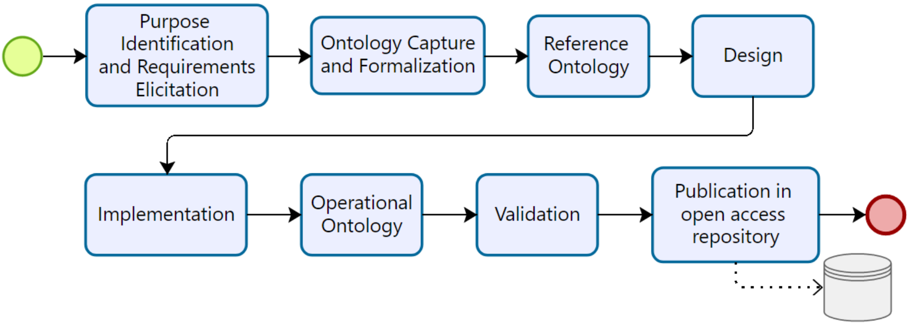

Purpose Identification and Requirements Elicitation
Ontology Capture and Formalization
Reference Ontology
Design
Implementation
Operational Ontology
Validation
Publication in open access repository

Figure 2.5 Creating Controlled Vocabulary as Ontology Sub-Process (Based on Falbo [73]).

The SABIO methodology [73] for ontology development consists of several key activities:

- Purpose Identification and Requirements Elicitation: This initial step involves defining the scope and purpose of the ontology. It includes gathering requirements from domain experts to ensure the ontology meets the necessary specifications and covers relevant concepts [73].
- Ontology Capture and Formalization: The identified requirements are translated into a formal ontology structure in this step. Key

concepts, relationships, and constraints are defined and captured in a structured manner [73].

- A reference ontology is developed to provide a high-level model representing the main concepts and relationships within the domain. This phase ensures the ontology is comprehensive and logically consistent [73].
- Design: The design phase involves detailing the structure and elements of the ontology, ensuring it aligns with the reference ontology and meets the specified requirements [73].
- Ontology Implementation: The formalized and designed ontology is implemented using an appropriate ontology language, such as OWL. This implementation phase converts the conceptual model into a machine-readable format [73].
- Operational Ontology: The implemented ontology is then deployed in an operational environment. This phase ensures that the ontology can be used effectively within the intended applications and systems [73].
- Validation: The ontology undergoes a validation activity to verify its accuracy and relevance. This activity involves testing the ontology with real-world data and feedback from domain experts to ensure it meets all requirements and performs as expected [73].

Finally, the resulting ontology should also be published in an open-access repository, just like the metadata schema.

2.5.4 Controlled Vocabulary Mapping

After identifying or creating a suitable ontology, the natural language values in the datasets should be replaced with the corresponding terms from the controlled vocabulary. Following this, the annotated dataset should be exported in a format that facilitates the next activities of data integration.

2.6 Validation

The validation of semantic interoperability (Fig. 2.6), i.e., ensuring that the data semantics are preserved, occurs in the final stage of the framework. At this point, datasets from different sources have been harmonized and annotated to a single data model. The annotated datasets from the previous step should align with the technologies chosen by the project team for integration or federated query. In some cases, simple CSV files may suffice, but more complex data structures, such as a RDF triple store, might be necessary.

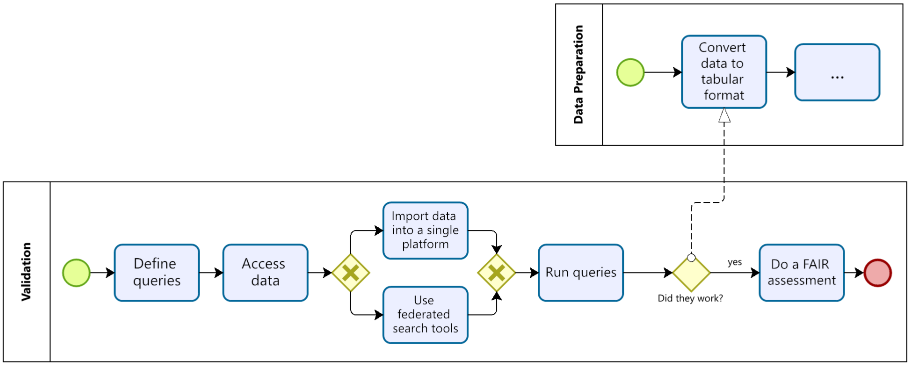


Figure 2.6 Validation Process.

To evaluate this semantic alignment, queries are executed to verify that the intended semantics are preserved and retrievable. In this context, these queries serve as operational indicators of the construct ”semantic interoperability” – they allow the validation of whether the conceptual meaning encoded in the annotations is properly realized at the technical level. Drawing from the notion of construct validity [92], this step aims to ensure that there is a meaningful and justifiable connection between the abstract concept (semantic alignment) and its practical measurement (query results).

If queries fail to return the expected results, it may suggest either an underrepresentation of the semantic model (i.e., some relevant aspects were not annotated) or construct-representation bias (i.e., the data was misannotated or transformed in a way that distorts meaning). Following Sjøberg and Bergersen’s typology [92], these are key threats to the construct validity of semantic models. Given the complexity of the framework and the opacity of error propagation, it is often challenging to isolate the specific source of semantic mismatch. In such cases, a pragmatic approach is to return to the Data Preparation phase to re-examine and, if needed, reconfigure earlier steps. The validation process should include a range of representative queries that reflect the intended uses of the data by the target user community. Adjustments to annotations or transformation pipelines may be necessary to improve semantic consistency and alignment.

Finally, a FAIR assessment should be conducted to ensure semantic interoperability and broader adherence to good data management practices. Although the FAIR principles extend beyond semantics, they offer a useful lens for evaluating the reusability and integration-readiness of the dataset. In Parts II and III of this dissertation, we present two case studies in the domains of agriculture and agricultural biodiversity, where we applied this framework and discuss its results.

CHAPTER 3

A Conceptual Model for FAIR Metadata Schemas

Chapter Overview

This chapter explores the design and creation of metadata schemas based on the FAIR Data Principles. We provide a clear interpretation of these principles, focusing on how they apply to metadata schemas. Leveraging the OntoUML language, we developed a conceptual model that explains the key components of a FAIR-compliant metadata schema. Through detailed discussion and provision of examples for each model component, this work aims to help metadata designers and curators better understand how to incorporate the FAIR Data principles into their schemas.

Publication note: This chapter is based on the publication of Soares et al. [93]: *Towards a Conceptual Model for FAIR Metadata Schemas*.

3.1 Introduction

Metadata schemas are tools that provide a framework for creating consistent metadata records. Essentially, these schemas define data structures and represent the semantic model of specific domains. Ideally, their development should be community-driven to incorporate diverse insights and ensure broad applicability.

Over recent decades, many communities have developed a wide range of metadata schemas, significantly contributing to data standardization across domains. One of the goals of employing these schemas is to enhance data's findability, accessibility, interoperability, and reusability, in line with the FAIR principles. However, many metadata schemas do not adhere to the FAIR Data principles themselves, making their reuse difficult [91, 94, 95]. To solve this, schema curators should apply the

FAIR principles in the development and management of these schemas. As outlined in FAIR principle I2, each community should decide the level of FAIRness required for the vocabularies they use. At a minimum, these vocabularies should have unique and permanent identifiers (F1), be accessible using standard methods (A1), and be described in a clear, widely understood language for knowledge representation (I1) [9].

This chapter explores the application of FAIR Data principles to the design and management of metadata schemas. This research has been inspired by our work on agriculture and biodiversity data management. We propose an OntoUML conceptual model that outlines key components essential for metadata schemas to comply with these principles.

3.2 Related Work

The application of FAIR data principles to semantic artifact curation is an evolving field, crucial for ensuring that resources like controlled vocabularies, taxonomies, ontologies, and metadata schemas support data FAIRification. Several studies have explored this area.

Berg-Cross and Arbor [96] highlight the importance of ontologies in making metadata computer-processable by providing axiomatized definitions. However, adapting FAIR principles to these resources introduces challenges, such as the need for community standards and meaningful semantics.

The FAIRsFAIR project report [94] proposes refined principles to improve the FAIRness of semantic artifacts, building on existing community recommendations. Poveda-Villalón et al. [95] examine the challenges of aligning ontologies with FAIR principles, emphasizing the need for ontologies that support data interoperability and reuse. Amdouni and Jonquet [91] present a framework for evaluating the FAIRness of semantic artifacts, offering a quantitative approach to assess adherence to FAIR principles. These efforts underscore the importance of establishing robust standards and practices for FAIR semantic artifacts construction, however none of them target metadata schemas specifically, neither present a conceptual model to elucidate the requirements for FAIR metadata schemas. Considering the importance of community metadata schemas for FAIR data, this chapter proposes a series of recommendations to make this specific type of semantic artifact FAIR.

3.3 A Brief Introduction to OntoUML Stereotypes

OntoUML is a language for specifing ontologies that has been implemented as a UML extension. Grounded in UFO, OntoUML provides a robust conceptual framework, making it an ideal tool for accurately representing complex domains [33, 34], such as in this chapter. UFO's comprehensive approach to ontology modeling, which includes key elements of real-world phenomena such as objects, events, and their interrelations, forms a solid basis for OntoUML. Consequently, OntoUML leverages this foundational strength to create 'detailed and precise' syntactic models; however, semantics remain within the model users' agreement. [33, 34].

OntoUML introduces several stereotypes to capture different kinds of ontological distinctions. These stereotypes offer a clear and precise way to model various elements within a domain, including, but not limited to: `Kind`, a natural category that provides identity to its instances; `Role`, an entity's role in a specific context; `Subkind`, a specialized kind that inherits properties from a broader kind while introducing unique features; `Mixin`, a characteristic shared by entities from different kinds; `RoleMixin`, which combines the concepts of role and mixin to represent a role that can be assumed by entities from different kinds; `Category`, an abstract grouping of mixins that does not provide identity; `Relator`, which mediates relationships between two or more entities; and `Quality`, a property or characteristic that an entity possesses, which can be measured or perceived [33].

OntoUML is particularly useful in scenarios where precise and unambiguous domain representations are crucial. For instance, in the development of FAIR metadata schemas, OntoUML can help create models that allows us to understand all the components and their relationships to define a FAIR metadata schema.

We defined our ontological conceptual model by following an approach similar to that proposed by Guizzardi [97], known as "ontological unpacking". This method involves analyzing a domain to reveal its underlying ontological conceptual model. Collaboration with specialists provided valuable insights that shaped this model. The following steps were undertaken:

Domain Analysis: This phase focused on identifying key concepts, entities, and relationships within the domain through structured meetings, workshops, and communications. Key discussions occurred during workshops promoted by the GO FAIR Agro Brazil Network[1], Brazil's 5th National Action Plan on Open Government[2], and the development of Almes Core [41]. Additionally, collaboration with the TDWG Biological Interaction Data Interest Group[3] during the development of the PPI Vocabulary [77] played a crucial role. Insights from the WorldFAIR Case Study on Agricultural Biodiversity [79, 80] also helped us to understand the key components of FAIR community metadata schemas.

---

[1] `https://go-fair-agro.github.io/`

[2] `https://www.opengovpartnership.org/wp-content/uploads/2022/01/Brazil_Action-Plan_2021-2023_EN.pdf`

[3] `https://github.com/tdwg/interaction`

Literature Consultation: Literature suggestions from domain experts and related works were also considered. The work of Marcia L. Zeng and collaborators [98, 99], and the ISO/IEC 11179 standards [100, 101, 102, 103] were key references for several concepts included in this model, due to their solid foundation for metadata schema modeling. The domain analysis combined with insights from the literature resulted in the requirements listed in Table 3.1.

Ontological Analysis: We examined the concepts in Table 3.1 to determine their ontological nature, and then classified these concepts into appropriate ontological categories, such as kinds, roles, phases, and mixins, in accordance with OntoUML.

Model Development: The conceptual model was developed based on the insights gained from the ontological analysis. This model represents the ontological structure of the domain of FAIR metadata schemas, capturing the essential nature and relationships of its concepts.

Validation: We validated the model by conducting a FAIR assessment grounded in the concepts outlined in this conceptual model.

Table 3.1 Adaptation of the FAIR Data Principles [8, 9] for Metadata Schema Management (Based on [94, 98]).

FAIR P.	Ref. P.	Refinement explanation
F1	RF1	Each metadata schema should have a unique identifier (i.e., the schema namespace), represented by a Uniform Resource Identifier (URI). Different versions of a metadata schema should have distinct identifiers.
F2	RF2	The landing page of a metadata schema should contain a descriptive metadata record with details about the schema. This metadata record should be machine-readable to allow search engines and other automated tools to discover the schema [104].
F4	RF3	To improve metadata schema findability, the schema and its metadata record should be indexed in vocabulary catalogs such as the Linked Open Vocabularies (LOV) [105], OntoBee [106], and the Digital Curation Centre List of Metadata Standards (`https://www.dcc.ac.uk/guidance/metadata-standards`).
A1, A1.1	RF4	The URIs implemented by the metadata schema should resolve to web pages accessible through widely used, universally accessible and secure communication protocols, such as HTTPS.
A2	RF5	The sustained accessibility of metadata schema documentation should be ensured, regardless of whether the schema is still active or maintained. This is important for applications that use past schema versions.
I1	RF6	Formal languages (e.g., RDF, OWL) should be used for the machine-readable representation of metadata schemas.

Table 3.1 (continued)

FAIR P.	Ref. P.	Refinement explanation
I1	RF7	Conceptual models in UML, OntoUML, or other languages should be provided to help humans understand the relationships between concepts.
I2	RF8	The range (data types) and the domain (a class) of datatype properties should be clearly specified in the schema.
I2, R1	RF9	The range (a class) and the domain (a class) for object properties should be clearly specified. If the classes used as range are defined elsewhere, this external controlled vocabulary should also adhere to the FAIR principles.
I2	RF10	Properties and classes should have clear definitions and names.
R1.1	RF11	Metadata schemas should be published under open licenses such as CC BY 4.0, and the license information should be machine-readable.
R1.2	RF12	Schema modifications, such as reviewing, deleting, and adding new terms should be documented.
R1.2	RF13	The authority responsible for curating the metadata schema should be indicated.
R1.3, I3	RF14	When developing new metadata schemas, reusing elements from existing schemas should be prioritized to minimize duplication. Reuse is recommended only when the elements have the same meaning, regardless of the schema context. The namespaces of all reused terms should be listed on the schema's landing page and included in its serializations.

3.4 Conceptual Model

Drawing on the refined FAIR Data principles in Table 3.1, supplemented by insights from the literature and the collective domain expertise of this chapter collaborators (refer to [93]), we present a conceptual model for FAIR metadata schemas in Fig. 3.1. We further explain each component of the conceptual model, providing examples.

3.4.1 Schema Authority

A FAIR metadata schema should be created through the collaborative efforts of specific communities of practice, to ensure they encode their discipline-specific elements [107]. Examples of community-developed schemas include DwC [47], Astronomy Visualization Metadata (AVM) [108], and the Data Documentation Initiative (DDI) [109]. Expert groups or institutions, often referred to as the schema `authority` [100, 102, 103, 110], should lead the development and curation of metadata schemas. The authority is responsible for any `modification` in the schema, which includes adding, reviewing, and removing metadata elements.

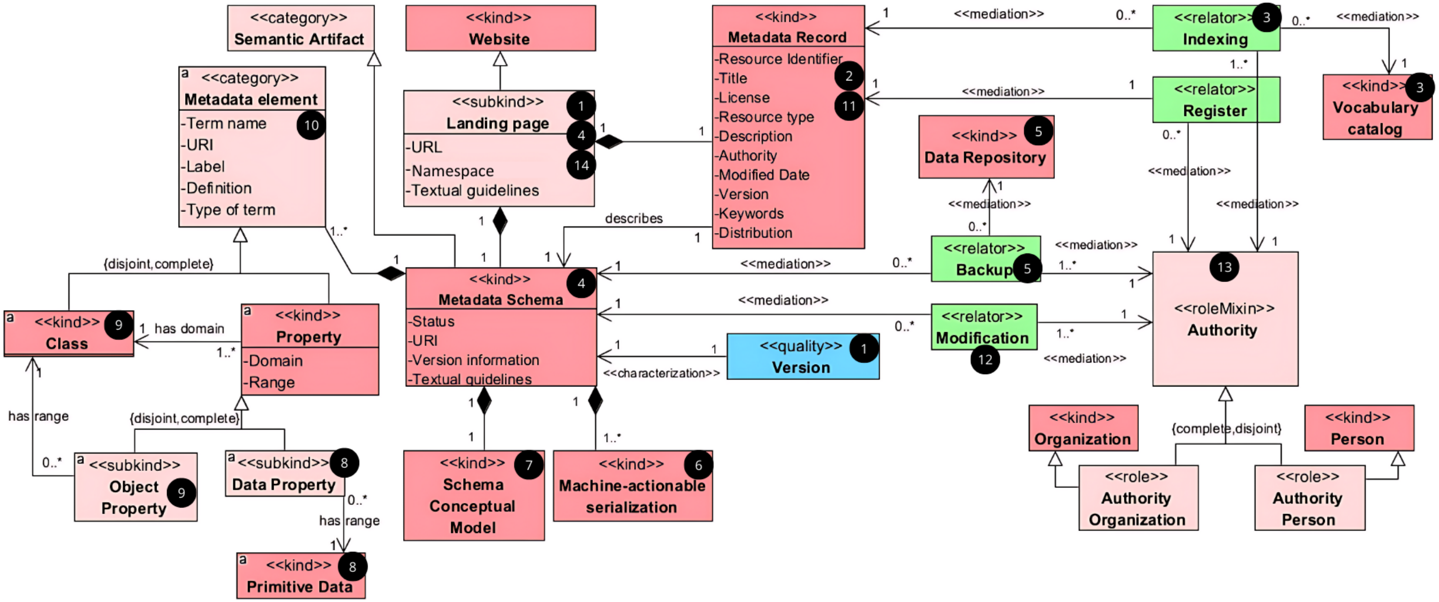


Figure 3.1 Conceptual Model of a FAIR Metadata Schema. Black circles represent refined FAIR principles from Table 3.1.

The authority behind the metadata schema should be publicly declared on the schema's landing page and within the schema metadata record. This transparency ensures that the user community knows who to contact for assistance, inquiries, suggestions, or participation in the curation community. The authority is also responsible for `indexing` the schema (and its descriptive metadata record) in vocabulary catalogs like LOV and OntoBee, so more users can find and reuse it.

Additionally, maintaining a `backup` of the metadata schema in data repositories like GitHub and Zenodo is crucial for its longevity, accessibility, and version control. Platforms like GitHub provide tools for tracking changes, reverting to previous versions, and facilitating effective collaboration.

When setting up a backup on GitHub, including a README file that explains the schema's purpose and provides guidelines for use and contribution is helpful. Utilizing GitHub features such as branches for ongoing developments, tags for stable releases, and releases for distributing specific versions can improve schema management. Engaging the community through issues and pull requests can also gather feedback and contributions, enriching the schema's development and adoption. Examples of community schemas using GitHub for this purpose include DwC[4], DDI-CDI[5], and Almes Core[6] [41].

3.4.2 Landing Page of a Metadata Schema

The `landing page` of a metadata schema is a type of `website` that consolidates all the information describing a schema. It typically includes details such as the schema's purpose, structure, and usage guidelines. It should be accessible via a URL that supports open data transfer protocols such as HTTPS, as recommended by RF5. This page should include at least the following components:

- A `schema metadata record`, describing the schema in formats accessible to both humans and machines.
- The `metadata schema` itself, providing the structure and specifications.

Furthermore, the landing page should declare the namespaces employed by the metadata schema, ensuring consistent terms usage. This includes declaring namespaces for terms reused from other schemas as well. For instance, DCAT defines its namespace as `http://www.w3.org/ns/dcat#`, reuses many terms from Dublin Core (DC), and introduces a minimal set of its own classes and properties[7].

---

[4] `https://github.com/tdwg/dwc`

[5] `https://github.com/ddialliance`

[6] `https://github.com/AlmesCore/`

[7] `https://www.w3.org/TR/2023/WD-vocab-dcat-3-20230307/`

3.4.3 Schema's Metadata Record

A schema `metadata record` typically appears as the first section of a metadata schema's landing page, as in the W3C standards, for instance. It should provide a comprehensive description accessible to both humans and machines. This record should include, at a minimum, the attributes listed in Table 3.2. This aligns with the recommendations for metadata discovery by [104], which provide guidelines on creating metadata records that can be efficiently discovered by search engines and other machines. CDIF [104] advises on using metadata terms from established schemas like DCAT, Schema.org, and Dublin Core to create the metadata record, and recommends using formats like JSON-LD for machine readability. The metadata record should be created (or `registered`) by the schema's authority.

Table 3.2 Attributes for a Schema's Metadata Record (Based on CDIF [104]).

Attribute	Description
Resource Identifier	A unique identifier for the metadata schema, such as its namespace.
Title	The official name of the metadata schema.
Distribution	URLs providing access to the schema, including its various representations (e.g., landing page, RDF/XML serialization).
Rights	The license under which the schema can be reused, with a preference for open licenses that offer machine-readable descriptions, like CC BY 4.0.
Resource Type	A label with a classification scheme detailing the kind of resource the metadata describes [104].
Description	A clear description of the schema, including its objectives and the domain it covers.
Authority	The person or organization responsible for the schema's curation, including at least a name and, if possible, contact information.
Modified Date	The date of the last modification to the metadata schema, formatted as "year" or in ISO 8601 date and time format, or any other machine-readable date format [104].
Version	The version label for the schema, if applicable. Version labels should allow alphanumeric sorting that reflects the sequence of version releases [104].
Keywords	Tags describing the metadata schema.

3.4.4 Metadata Schema

A `metadata schema` is a type of `semantic artifact` aimed at organizing, categorizing, and managing information and knowledge within a specific domain or context [111, 112]. A metadata schema requires comprehensive documentation to ensure it is accessible and reusable for humans and computer applications alike. This documentation comprehends several components, as detailed in Fig. 3.1. A key component of metadata schemas is its namespace, which distinguishes it from other schemas.

Metadata schemas have as basic components `metadata elements` and the relationships between these elements [98]. In an ontological approach, these elements are usually classified as `classes` or `properties` [98, 113]. They are typically described in textual form (as a glossary of terms) on the schema's landing page for human consumption. Additionally, these schemas should be represented in both conceptual and machine-actionable models.

3.4.5 Metadata Element

A `metadata element` can be either a `class` or a `property` [98]. A class represents a category or type of thing (e.g., Book, Person, Organism) and is a way to group similar items based on shared characteristics [114]. Classes provide a high-level organizational structure to metadata. They allow for categorizing different entities or concepts within the described domain. For instance, the DwC metadata schema (developed for the biodiversity domain) includes classes such as `dwc:Event`, `dwc:Occurrence`, and `dwc:Taxon`.

Properties describe specific characteristics or aspects of an individual within a given class. For the class `dwc:Event`, examples of properties include `dwc:eventDate`, `dwc:eventID`, and `dwc:EventMeasurement`. Properties in metadata schemas, as in ontologies, can be classified into two types: `object property` and `data property`.

Object properties define relationships between two entities (or classes) within the metadata schema. They link instances of one class to instances of another class [115]. For example, in a bibliographic metadata schema:

```
:author a owl:ObjectProperty ;
        rdfs:domain :Book ;
        rdfs:range :Person .
```

In this example, `:author` is an object property that relates a `:Book` to a `:Person`. This indicates that a book can have one or more authors, and each author is a person.

Data properties (also known as datatype properties) define relationships between an entity and primitive data types. They link instances of a class to values such as strings, numbers, or dates (i.e., `primitive data types`) [115, 116]. For example, in the bibliographic schema:

```
:publicationDate a owl:DatatypeProperty ;
                 rdfs:domain :Book ;
```

```
rdfs:range xsd:date .
```

The domain of a property specifies the class of individuals to which the property can be applied. In other words, if a property has a certain class as its domain, then any individual that possesses this property is assumed to be a member of that class. For instance, if a property `:hasAuthor` has a as domain `:Book`, then any resource that `:hasAuthor` is inferred to be a Book [115].

Defining domains and ranges between concepts in a metadata schema enhances the precision and clarity of the relationships between data elements and enables machine inferences [117]. By specifying which classes of objects a property can relate to (domain) and the type of values it can take (range), we infer that data elements belong to the specified classes.

In addition to the relationships between elements in a schema, each metadata element should include a specification of semantics and content [98, 99, 100]. Semantics is about the “meanings of the elements”, and content is related to providing “general instructions on what and how values should be assigned to the elements in an application” [98, p.39].

3.4.6 Element Semantic Specification

Unclear semantic specification of the elements in a metadata schema hamper its reuse. A combination of attributes can be used to define the semantics of an element. For instance in DC, ‘term name’ assigns a specific, distinct name to the metadata element within the schema. A ‘Label’ is included for human readability and can be available in multiple languages. ‘Definition’ provides in-depth explanations regarding the element’s purpose, scope and use.

The ISO/IEC 11179 standards, grounded on the Theory of Terminology, guide articulating the semantics of metadata elements, focusing on formulating definitions in Part 4 [102] and term names in Part 5 [103].

Formulating Metadata Definitions

Part 4 of ISO/IEC 11179 outlines a set of requirements and recommendations for metadata definitions [102]. The requirements specify that definitions should focus on singular concepts, express the essence of the concept rather than what it is not, and be formulated as descriptive phrases or sentences. They should also avoid uncommon abbreviations and not embed definitions of other data or underlying concepts. Additionally, the recommendations advise that definitions convey the essential meaning of the concept, be precise, unambiguous, concise, and able to stand alone. Definitions should also avoid embedding rationale, functional usage, domain information, or procedural details. Circular reasoning should be avoided, and consistent terminology and logical structure should be maintained for related definitions.

Formulating Metadata Names

In addition to providing clear definitions, metadata elements must have appropriate names. ISO/IEC 11179-5 [103] provides naming conventions to ensure names are meaningful and easily understood by users. In this standard, each name represents a concept in a metadata schema and follows specific conventions, which include syntactic, semantic, and lexical rules that should be documented. The guiding principles for naming include defining the scope (general or specific), identifying the naming authority, specifying the components (semantic principle), arranging them (syntactic principle), and determining their appearance (lexical principle). Additionally, the uniqueness principle ensures names are distinct within the defined scope, with differentiation rules if needed.

ISO/IEC 11179-5 provides a flexible framework for nomenclature, adaptable across domains. The semantics of name parts depend on the type of metadata term. While class-type terms may be abstract, they must still be named and hierarchically organized within the schema, similar to other metadata terms, in relation to their properties [103].

Property name consistency follows two main rules: **Discreteness**, where each property term has a unique meaning without overlap, and **Completeness**, ensuring that the set of property terms fully describes an entity [103]. Property names can indicate value domains, such as Color, Quantity, or Height, as seen in the PPI schema, with terms like `ppi:LengthInMilimiters`, `ppi:ovuleQuantity`, and `ppi:visitedFlowersQuantity` [77].

A property name may also include a qualifier term [103]. A qualifier can be used to distinguish metadata terms with identical names or to represent more specific values, such as the qualifiers ‘start’ and ‘end’ in `dcat:startDate` and `dcat:endDate`.

Finally, naming conventions should include the use of separators like hyphens, grouping symbols, or formatting resources like CamelCase, to connect name parts, for both classes and properties [103].

Machine-Actionable Serialization

Semantic Web languages like RDF and OWL allow the creation of machine-readable and actionable serializations of metadata schemas. Initially, early metadata schemas, including DC and VRA Core, were designed as purely XML schemas. Over time, particularly in the second decade of the millennium, there was a shift towards RDF vocabularies [98].

FAIR metadata schemas should be implemented as Semantic Web schemas, using languages such as RDF and OWL, as outlined in principle RF7 in Table 3.1. These languages enable advanced functions like semantic querying and automated reasoning. RDF provides a framework for expressing information about resources in a graph form, which is highly beneficial for linking data [118]. OWL adds a richer ontology layer, representing more complex relationships and classifications, enhancing the depth and utility of the schema [119]. The choice of serialization language depends on the goals governing authority of the schema. A persistent identifier (URI or IRI) should be used as the schema namespace. Each metadata element should then have a unique identifier within that namespace. There are services that help generate these identifiers, such as the W3C Permanent Identifier Community Group[8] and purl.org[9].

Schema Conceptual Model

A conceptual model is pivotal in developing and implementing a metadata schema. It is an abstract representation of the phenomena the metadata schema aims to describe, capturing the essential concepts and their relationships, sometimes referred to as metamodel or upper model. A conceptual model may serve different purposes in a metadata schema:

- Schema design: A conceptual model helps identify the key concepts, entities, and attributes relevant to the described domain, ensuring that the schema covers all necessary aspects of the data it intends to represent [67].
- Data integration and mapping: Conceptual models may provide a reference point for mapping between schemas. They are a fundamental component for enabling semantic interoperability between heterogeneous datasets [67].
- Human understanding: Conceptual models are useful for schema designers and end-users. They can facilitate the understanding of the schema by providing a visual, conceptual representation of its classes and properties. This also makes the schema easier to reuse by users unfamiliar with it [67].

Many metadata schemas use conceptual models for these purposes. UML is one of the most popular language used for metadata schema conceptual models. For instance, EML[10] is based on a UML conceptual model that defines the structure and semantics of metadata to describe ecological data effectively. Besides EML, the DDI-CDI Framework is

---

[8] `https://w3id.org/`

[9] `https://purl.org/`

[10] `https://eml.ecoinformatics.org`

another example of a schema that also uses a UML conceptual model, which is described as the "core" of their schema[11].

3.5 Final Considerations

This chapter underscores the importance of aligning metadata schemas with the FAIR Data Principles, serving as a guide for curators and designers. The interpretations of these principles may vary, and the guidelines provided here are meant to support and inform rather than prescribe specific actions.

In conclusion, creating and managing metadata schemas is a nuanced and intricate endeavor, requiring thoughtful integration of various factors and standards. Commitment to best practices, such as prioritizing clarity, reusability, and interoperability, facilitates the development of metadata schemas that effectively meet diverse data management needs. The perspectives and recommendations offered in this study aim to aid in creating FAIR-compliant metadata schemas, thereby enriching the data management and sharing ecosystem. As metadata schemas align more closely with FAIR principles, the management and utilization of digital objects will significantly improve, enhancing both accessibility and utility.

---

[11] `https://ddialliance.org/Specification/ddi-cdi#uml_model`

Part II

Case Study I: Harmonization and Integration of Agricultural Price Index Datasets

Introduction to Part II

Agriculture and agricultural biodiversity have been chosen as the two fields to implement our semantic interoperability framework. Within these fields, specific domains were selected for detailed analysis, and a collection of data sets was chosen for in-depth study. Case Study 1 describes the initial implementation of the framework, while Case Study 2 was developed to test the framework in a different domain, demonstrating its applicability across multiple domains.

The first domain is agricultural trading, which encompasses the market prices used in the trading of agricultural commodities. This domain is critical for understanding market dynamics, price fluctuations, and the economic factors influencing agricultural production and distribution [120]. Accurate and timely data on agricultural trading prices is essential for farmers, traders, and policymakers to make informed decisions and manage risks effectively [121]. For instance, real-time price data can help farmers decide when to sell their produce to maximize profits or minimize losses [120].

CHAPTER 4

Goal Definition and Data Preparation

Chapter Overview

This chapter outlines the initial steps of Case Study 1, aimed at creating a unified framework for publishing agricultural price index data, driven by the Brazilian 5th National Action Plan on Open Government. Through collaboration with key Brazilian institutions like Embrapa, CEPEA, Ipea, and CONAB, we addressed fragmented data practices in agriculture to enhance FAIR-aligned data integration. Beginning with Section 4.1, the chapter describes our goal of improving data interoperability and accessibility for stakeholders. Section 4.2 then provides an overview of datasets from CEPEA, Ipea, and CONAB, noting their unique structures, scope, and contributions to agricultural data. In Section 4.3, a standardized template for metadata and data integration is introduced, defining core elements like product identifiers, descriptions, and publication details. Lastly, Section 4.4 details the cleaning and formatting processes, using tools like OpenRefine and Python scripts to create a consistent dataset for further semantic processing.

4.1 Goal Definition

The goal for Case Study 1 was defined through the process outlined in Section 2.3 by a multi-stakeholder group representing key institutions across Brazil. The objective was to design a common framework for publishing price index data to enable data convergence into a single integrated platform. This initiative was driven by the 5th Brazilian National Action Plan on Open Government, launched in 2021 as part of Brazil's commitment to the international Open Government Partnership [18].

The 5th Brazilian Action Plan for Open Government included a commitment to promote the opening and integration of agricultural value chain databases in an approach that meets public interest [18]. The action was led by the Brazilian Agricultural Research Corporation (Embrapa) to solve the problem of having fragmented efforts on data opening among the numerous actors in the Brazilian agricultural sector. Even though many initiatives have been launched to open agricultural data, they often work in fragments and inconsistently, thus limiting the usability and access of the data [18].

The commitment addressed the lack of cohesiveness in data management and the non-adherence to open data principles [18]. Data was distributed and not properly treated, affecting it findability, accessibility, interoperability, and reusability (FAIR). Although systems have been developed by CONAB, Embrapa, and the Ministry of Agriculture, Livestock, and Supply (Mapa) for data collection and disclosure, they still lacked integration and alignment with the needs of the society [18].

The commitment triggered a systematic effort to solve the collaboration problem between government bodies and society. Among the first steps was the organization of meetings to take stock of existing databases and to identify priority areas for data opening and integration [18]. This collaborative process was designed to align the work of many stakeholders, eliminate duplication, and fulfill the demands of civil society. After these assessments, the project appraised the risks and impacts of opening and integrating databases and their feasibility for ethical and responsible data reuse. It also defined the minimal requirements for metadata that should be available for interoperability and data transparency [18].

Following the launch of an action plan to fulfill the commitment, additional institutions joined the initiative. The GO-FAIR Agro Brazil Thematic Network [20, 76] has played a pivotal role in coordinating and connecting these institutions. This network brought together experts in FAIR Data Principles, metadata, data stewardship, and computer science, along with domain specialists from Brazilian institutions such as Embrapa, CEPEA, and C4AI.

C4AI, a collaborative initiative by the University of São Paulo (USP), the São Paulo Research Foundation (FAPESP), and the International Business Machines Corporation (IBM), aims to establish a leading research center in Artificial Intelligence (AI) in Brazil. The center is organized into eight research groups focused on AI applications across different domains. Among these is the AgriBio: Causal Multicriteria Decision Making in Food Production Networks Group, which has prioritized improving data quality in agriculture to support more effective AI processing. This PhD project has been developed as part of this effort.

After identifying all the partners and establishing the necessary agreements, we began integrating datasets from three Brazilian institutions that publish agricultural price index open data: CEPEA, Ipea, and CONAB. As part of this effort, this PhD project developed a prototype triple store, named C4AI-KGAP presented in Chapter 9, to integrate and manage the data from these institutions.

4.2 Selected Datasets

The datasets selected for this study were sourced from three Brazilian institutions that publish agricultural price index open data: CEPEA, Ipea, and CONAB. These institutions provide critical information for understanding agricultural market dynamics in Brazil, serving stakeholders such as policymakers, producers, researchers, and traders. Integrating data from these sources aimed to address the fragmentation and inconsistency in data management, a key goal of the broader initiative described in Section 4.1.

Each dataset represents unique characteristics in terms of structure, scope, and granularity. The CEPEA dataset is highly detailed and updated daily, offering extensive coverage of agricultural and livestock products. The Ipea dataset aggregates products into broader categories and provides monthly price indexes, emphasizing market trends rather than individual commodities. The CONAB dataset, the most comprehensive among the three, includes weekly and monthly prices for over 130 agricultural products across all Brazilian regions. The integration of these datasets required significant preprocessing and metadata enrichment to standardize their formats and ensure semantic interoperability.

4.2.1 Cepea Database

The CEPEA Database is extensive, offering a broad range of data on agricultural and livestock products. It includes open-access information[1] on commodities such as cattle, coffee, and cotton, among others [122]. CEPEA publishes daily price indexes for these products [122]. The database allows users to apply filters and download the data as an `.xls` dataset[2] containing raw data, as illustrated in Table 4.1.

In addition to the basic price indexes, the CEPEA website includes detailed reports and analyses, providing a robust foundation for understanding the dynamics of agricultural markets in Brazil [122]. However, this additional information is only available on their website and it is

---

[1] The CEPEA database has also a private subscription, which includes more products than the open access version. However, for this research, only open access data were used, in line with our commitment to the Open Access Movement, as previously discussed in this dissertation.

[2] In the literature, the terms ‘Database’ and ‘Dataset’ are often used interchangeably. However, in this dissertation, database refers to the entire system where institutions store and manage their data, while dataset specifically refers to the subsets or snippets of data extracted from these databases for analysis in this study.

not included in the downloaded datasets. This data is essential for farmers, traders, policymakers, and researchers who require accurate and timely information to make informed decisions and develop effective strategies in the agricultural sector [122]. CEPEA's indexes are widely used for financial settlement, reflecting the actual product delivered. Even individuals unfamiliar with the stock market rely on CEPEA's indexes for physical market transactions (buying and selling outside the stock exchange, potentially mediated by trading offices). Supermarket chains and other retail networks also utilize CEPEA's indexes. Major news agencies such as Reuters and Bloomberg use CEPEA's indexes for their reporting and analyses. This widespread usage underscores the importance and reliability of CEPEA's indexes in various agricultural market segments [122].

Table 4.1 Snippet of the CEPEA Database in Portuguese [123].

INDICADOR DO AÇÚCAR CRISTAL BRANCO CEPEA/ESALQ - SÃO PAULO Fonte: CEPEA		
Data	**À vista R$**	**À vista US$**
21/12/2004	30,85	11,42
22/12/2004	30,91	11,42
23/12/2004	30,49	11,28
27/12/2004	30,4	11,3
28/12/2004	30,27	11,25
29/12/2004	30,25	11,34
30/12/2004	29,9	11,26

In Table 4.1, the top row is the title of the data snippet, which combines three elements: the product type 'Açúcar Cristal Branco' (Crystal Sugar), the price index creator (CEPEA/ESALQ), and the location (São Paulo). The 'Fonte' (source) element indicates the provenance of the data, in this case CEPEA. 'Data' (date) indicates, for example, that on December 21, 2004, the price of crystal sugar was 30.85 Brazilian Reais or 11.42 U.S. dollars.

The original price index is published in Brazilian reais (R$), and the conversion to other currencies, such as US dollars, is a modification of the original data. For this reason, in this case study, we only considered prices in Brazilian reais.

In this case study, we sampled the CEPEA Database by filtering data for Crystal Sugar in the São Paulo and Santos regions, as well as Cattle, Arabica Coffee, Robusta Coffee, and Soybean in the Paranaguá and Paraná regions. A separate dataset was generated for each product type, and all datasets are publicly available on Zenodo [1].

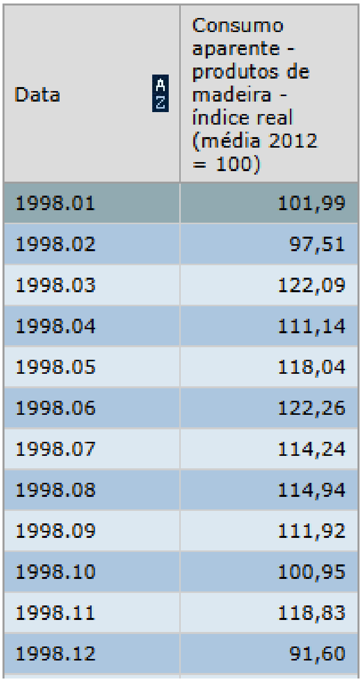
4.2.2 Ipea Database

The IpeaData database provides data on diverse products beyond the agricultural sector. As for agricultural products, it publishes price index data for the following categories:

- leather and leather goods;
- cellulose pulp, paper, and paper products;
- tobacco products;
- wood products.

The Ipea database, also called just IpeaData, provides prices for aggregations of products rather than providing individual product price indexes [124]. The price indexes are published monthly and are valuable for understanding broader market trends rather than the specifics of individual commodities. IpeaData allows to download data subsets as `.xls` files or other formats. Table 4.2 shows an example of the structure of a dataset downloaded from IpeaData.

Table 4.2 Snippet of the IpeaData Database in Portuguese [125].



Data	Consumo aparente - produtos de madeira - índice real (média 2012 = 100)
1998.01	101,99
1998.02	97,51
1998.03	122,09
1998.04	111,14
1998.05	118,04
1998.06	122,26
1998.07	114,24
1998.08	114,94
1998.09	111,92
1998.10	100,95
1998.11	118,83
1998.12	91,60

Table 4.2 shows a dataset with only two columns. The first column represents the publication date of the price index, while the second column first row contains a lengthy label that includes the type of product (produtos de madeira), type of statistics (índice real – média

2012 = 100), and the name of the publishing institution (Instituto de Pesquisa Econômica Aplicada), which could be considered as metadata. However, none of this information describes the data associated with this column, which are the prices in Brazilian reais.

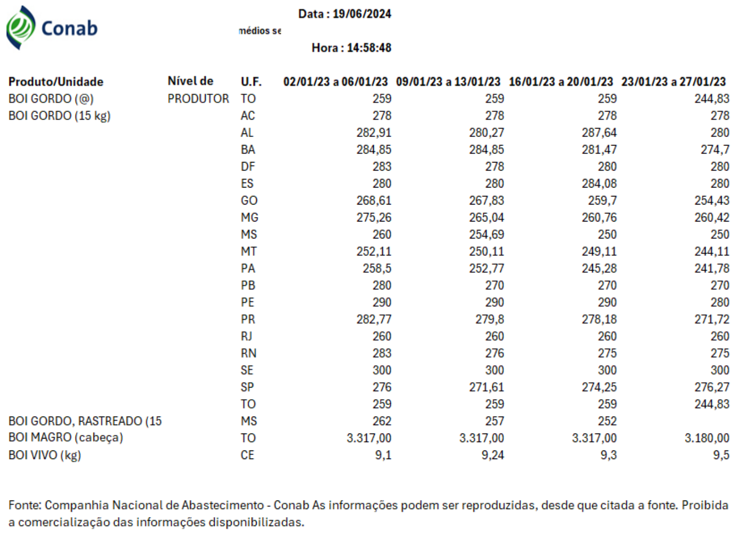
In this case study, we sampled the Ipea database by selecting data for leather, cellulose, paper, tobacco, and wood products. The data spans from the inception of these series to June 2023. No location filters were applied, as the prices published by Ipea represent nationwide data. A separate dataset was generated for each product type, and all datasets are publicly available on Zenodo [2].

4.2.3 CONAB Database

CONAB publishes the most extensive list of agricultural price indexes among the three data providers (in open access). For over 30 years, CONAB has collected price data for over 130 agricultural products, resulting in a database of more than 20,000 records. The data covers all regions of Brazil and is guided by public policies and demands. Price information is crucial for agricultural decisions, influencing crop choices for producers and purchase decisions for consumers [126].

CONAB publishes prices for these products on a weekly or monthly basis [126]. The data is available for download in `.xls`, HTML, or PDF formats. Users can generate customized datasets by applying filters based on the available data in the database. Table 4.3 shows a dataset with weekly prices, while Table 4.4 shows a dataset with monthly prices.

Table 4.3 CONAB Weekly Prices for Fed Cattle in Portuguese [127].



médios se

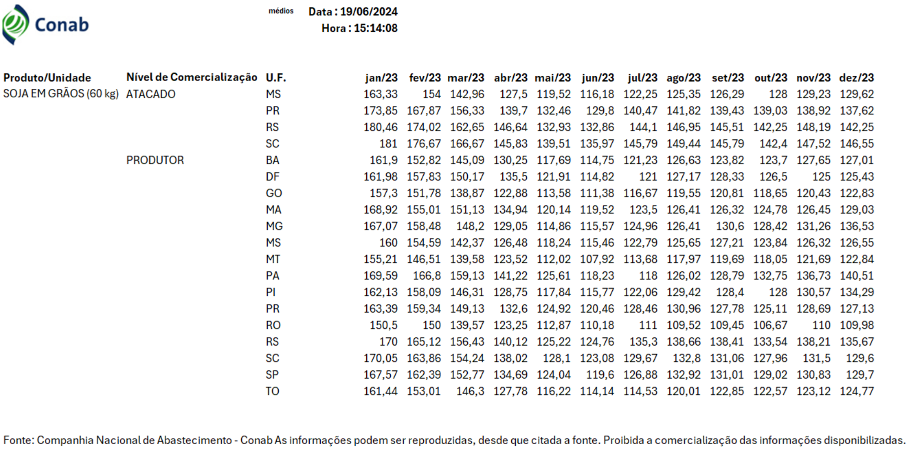
Data : 19/06/2024

Hora : 14:58:48

Produto/Unidade	Nível de	U.F.	02/01/23 a 06/01/23	09/01/23 a 13/01/23	16/01/23 a 20/01/23	23/01/23 a 27/01/23
BOI GORDO (@)	PRODUTOR	TO	259	259	259	244,83
BOI GORDO (15 kg)		AC	278	278	278	278
		AL	282,91	280,27	287,64	280
		BA	284,85	284,85	281,47	274,7
		DF	283	278	280	280
		ES	280	280	284,08	280
		GO	268,61	267,83	259,7	254,43
		MG	275,26	265,04	260,76	260,42
		MS	260	254,69	250	250
		MT	252,11	250,11	249,11	244,11
		PA	258,5	252,77	245,28	241,78
		PB	280	270	270	270
		PE	290	290	290	280
		PR	282,77	279,8	278,18	271,72
		RJ	260	260	260	260
		RN	283	276	275	275
		SE	300	300	300	300
		SP	276	271,61	274,25	276,27
		TO	259	259	259	244,83
BOI GORDO, RASTREADO (15		MS	262	257	252	
BOI MAGRO (cabeça)		TO	3.317,00	3.317,00	3.317,00	3.180,00
BOI VIVO (kg)		CE	9,1	9,24	9,3	9,5

Fonte: Companhia Nacional de Abastecimento - Conab As informações podem ser reproduzidas, desde que citada a fonte. Proibida a comercialização das informações disponibilizadas.

Table 4.4 CONAB Monthly Prices for Soybean in Portuguese [127].



Conab

médios Data : 19/06/2024
Hora : 15:14:08

Produto/Unidade	Nível de Comercialização	U.F.	jan/23	fev/23	mar/23	abr/23	mai/23	jun/23	jul/23	ago/23	set/23	out/23	nov/23	dez/23
SOJA EM GRÃOS (60 kg)	ATACADO	MS	163,33	154	142,96	127,5	119,52	116,18	122,25	125,35	126,29	128	129,23	129,62
		PR	173,85	167,87	156,33	139,7	132,46	129,8	140,47	141,82	139,43	139,03	138,92	137,62
		RS	180,46	174,02	162,65	146,64	132,93	132,86	144,1	146,95	145,51	142,25	148,19	142,25
		SC	181	176,67	166,67	145,83	139,51	135,97	145,79	149,44	145,79	142,4	147,52	146,55
	PRODUTOR	BA	161,9	152,82	145,09	130,25	117,69	114,75	121,23	126,63	123,82	123,7	127,65	127,01
		DF	161,98	157,83	150,17	135,5	121,91	114,82	121	127,17	128,33	126,5	125	125,43
		GO	157,3	151,78	138,87	122,88	113,58	111,38	116,67	119,55	120,81	118,65	120,43	122,83
		MA	168,92	155,01	151,13	134,94	120,14	119,52	123,5	126,41	126,32	124,78	126,45	129,03
		MG	167,07	158,48	148,2	129,05	114,86	115,57	124,96	126,41	130,6	128,42	131,26	136,53
		MS	160	154,59	142,37	126,48	118,24	115,46	122,79	125,65	127,21	123,84	126,32	126,55
		MT	155,21	146,51	139,58	123,52	112,02	107,92	113,68	117,97	119,69	118,05	121,69	122,84
		PA	169,59	166,8	159,13	141,22	125,61	118,23	118	126,02	128,79	132,75	136,73	140,51
		PI	162,13	158,09	146,31	128,75	117,84	115,77	122,06	129,42	128,4	128	130,57	134,29
		PR	163,39	159,34	149,13	132,6	124,92	120,46	128,46	130,96	127,78	125,11	128,69	127,13
		RO	150,5	150	139,57	123,25	112,87	110,18	111	109,52	109,45	106,67	110	109,98
		RS	170	165,12	156,43	140,12	125,22	124,76	135,3	138,66	138,41	133,54	138,21	135,67
		SC	170,05	163,86	154,24	138,02	128,1	123,08	129,67	132,8	131,06	127,96	131,5	129,6
		SP	167,57	162,39	152,77	134,69	124,04	119,6	126,88	132,92	131,01	129,02	130,83	129,7
		TO	161,44	153,01	146,3	127,78	116,22	114,14	114,53	120,01	122,85	122,57	123,12	124,77

Fonte: Companhia Nacional de Abastecimento - Conab As informações podem ser reproduzidas, desde que citada a fonte. Proibida a comercialização das informações disponibilizadas.

Both weekly and monthly datasets provide information on the type of product (fed cattle, soybean), level of trading, reference quantity (@, 15kg, 60kg), location, frequency, date and time of download, and the date of each price data point. In Table 4.3, the date column header represents a time interval marking the beginning and end of the week, while in Table 4.4, the date is provided in the month/year format.

In this case study, we sampled the CONAB Prices Database [126], selecting price indexes for Fed Cattle (weekly prices from January 2023 to August 2023), as well as Sugar, Coffee, and Soybean (monthly prices from January 2023 to December 2023). No location filters were applied, allowing us to download data for all regions available in the database. A separate dataset was generated for each product type, and all datasets are publicly available on Zenodo [3].

4.3 Creating the Data Template

The process for creating of a data template was described in Section 2.3. In this case study, we divided the template in two parts: one for the metadata and another for the data. During discussions with specialists, we identified key data concepts to include in the metadata template. These concepts are:

- **ID**: The Dataset ID. Since the data creators did not provide an ID for the datasets, we created artificial IDs to be used in the prototype.
- **Product group**: e.g., livestock, vegetables.
- **Product type**: e.g., fed cattle, soybean, sugar.

- **Title**: The title given to the dataset by the creator, such as "INDICADOR AÇÚCAR CRISTAL BRANCO CEPEA/ESALQ - SÃO PAULO."
- **Description**: Details of the data, including relevant information regarding statistical methods, tax incidence over the price, etc.
- **Creator**: The creator of the dataset.
- **Publisher**: The organization that published the dataset.
- **Reference**: The URL of the source from which the dataset was retrieved.
- **Descriptive statistics**: Free text describing the type of descriptive statistics used to calculate the prices.
- **Periodicity**: Frequency of the data publication (enumeration).
- **Reference quantity**: Given in Brazilian reais per kilo.
- **Location**: Name of the location related to the data.
- **Start date of the series**: The beginning date of the data series.
- **End date of the series**: The end date of the series, left blank if the series is still active.
- **License**: URL of the usage license.
- **Rights**: Rights associated with the dataset.
- **Methodology**: A URL of a document describing the methodology used to calculate the prices.

As for the data template, it included the following headers:

- **Metadata ID**: An foreign key from the metadata table.
- **Date of publication**: The date the price was published.
- **Price**: The amount of money.
- **Currency**: Normally Brazilian reais (R$).

These concepts were primarily derived from the data structures present in the datasets we analyzed, as well as suggestions from the specialists. This data template served as the basis for deciding which data to add or remove from the downloaded datasets as described in Section 4.4. It also formed the foundation for modeling Almes Core in a later phase of the project (refer to Section 5.2).

4.4 Data Preparation

In the data preparation process, as described in Section 2.4, we performed the initial processing of the datasets described in Section 4.2. This processing consisted mostly of filling up the template with data from the datasets generated in the previous step described in Section 4.2.

4.4.1 Filling Up the Template

The filled up templates with the metadata and data used in this case study are available on Zenodo in CSV format [128]. Since the datasets deployed in the previous stage lacked much of the information needed to complete the metadata template, we manually extracted most of this information from the institutions' websites.

The CEPEA and IpeaData websites provided all the necessary information to complete the metadata template. However, CONAB did not provide a title for the datasets, so we created artificial titles by combining the frequency (either "average monthly prices" or "average weekly prices"), product name, and commercialization level (e.g., wholesale, 'atacado' in Portuguese, for direct sales to the final customer, or "producer", 'produtor' in Portuguese, for the price paid to the farmer). Additionally, CONAB Datasets lacked descriptions, start dates for the price series, and links to the methodology used for price calculations. We requested the missing data from CONAB through the Brazilian Law No. 12.527/2011 [15], which ensures the constitutional right of access to public information. This request was submitted via the Gov.br portal[3]. CONAB provided incomplete information, and as result, these fields remain blank in the datasets, which limits their potential applications.

A single file was generated containing the metadata for the datasets from all three institutions (available as `metadata_template.csv` on Zenodo [5]).

After completing the metadata template, we populated the data template. This involved transferring only the relevant data from the original datasets into the standardized data template. Next, we used a Python script (Available on Zenodo [4]) to merge the multiple files generated in the previous step into a single dataset for each institution. An exception was made for CONAB, where separate datasets were created for weekly and monthly prices. The number of tuples in each resulting dataset is as follows:

- CEPEA: 36803 rows;
- Ipea: 1224 rows;
- CONAB Weekly prices: 768 rows;
- CONAB Monthly prices: 432 rows.

After integration, we proceeded to standardize data types in the data template using the software OpenRefine. It worked well in some cases, for instance for converting date values from the CEPEA Dataset, from the dd/mm/yy format to the yyyy-mm-dd format. However, when we attempted to convert the date types from the CONAB and IpeaData Datasets, which are presented with the first three letters of the month

---

[3] `https://www.gov.br/acessoainformacao/pt-br`

in English and the last two digits of the year (e.g., Jan-23), and as a combination of the year and the month separated by a dot (e.g., 1998.07), respectively, it did not work in OpenRefine. Consequently, we developed a Python script (available on Zenodo [6]) to perform these conversions.

Finally, we finished the process of data preparation and published a new version of the datasets, which is available on Zenodo [7]. We then moved forward to the next step in the semantic interoperability framework, which is the semantic annotation process.

CHAPTER 5

Semantic Annotation

Chapter Overview

This chapter details the semantic annotation process described in Section 2.5, which was conducted in two stages: metadata and data annotation. Given the lack of suitable Knowledge Organization System (KOS) to fully annotate agricultural price datasets, we developed a new schema Almes Core. It was designed to represent agricultural product price index data, with properties for product type, frequency, location, and methodology, drawing on controlled vocabularies and widely-used schemas like Dublin Core. Sections 5.2.1 and 5.2.7 describe the schema's structure. In Section 5.2.4 we discuss the challenges and solutions in assigning persistent URIs and establishing FAIR-compliant metadata schemas. Finally, Section 5.3 explains the process of aligning dataset values with Almes Core, using controlled vocabularies and customized fields where necessary. This mapping ensured coherent integration of agricultural data, leading to the development of a new ontology, Agricultural Products Type Ontology (APTO), to represent agriculture product types.

Publication note: Part of this chapter is based on the publication of Soares et al. [41]: *Building a Community-Based FAIR Metadata Schema for Brazilian Agriculture and Livestock Trading Data*.

5.1 Searching for Metadata Schemas

We began the semantic annotation process (as described in Section 2.5) by searching for vocabularies that could be reused in a new KOS. We conducted a bibliometric analysis of the literature, the results of which were detailed in the paper [129]. Although the bibliometric study provided insights into the main applications of KOS in agriculture, it did not assist in selecting the KOS to be reused in this case study. The recommendations for suitable KOS for reuse were made by specialists.

The first step in specifying this new metadata schema was gathering with specialists to define the schema's core elements, based on the purpose of the project, which was annotating price index datasets, and the data template used in the previous steps of the semantic interoperability framework. After that, the elements went through a second round of review before we released the first version of the element set, which was described in [41].

5.2 Creating a New Metadata Schema

Version 1 of Almes Core was a simple list of properties based on the templates presented in Section 4.1. This version was later revised and updated to meet the requirements for FAIR metadata schemas, as discussed in Chapter 3. New terms, including properties and classes, were added, some terms were removed, and the schema conceptual model was defined. Each term was assigned a unique name and definition, as well as a URI within the `https://w3id.org/AlmesCore#` namespace. This latest version of the schema (v. 2.0.1) modeled range and domain relationships between concepts, which guided its implementation in RDF.

Almes Core is available in two forms: Simple Almes Core and Extended Almes Core, both published on GitHub[1]. The Extended version includes additional classes specifically designed for use as RDF, and is defined under the `https://w3id.org/AlmesCore/rdf#` namespace. It introduces the Agent class, which replaces literal values for dc:creator and dc:publisher with structured entities, and the GeoNamesFeature class, which enhances sdo:location by linking to geographic entities.

In contrast, the Simple form remains compatible with RDF but allows creator, publisher, and location properties to accept literals as values. The Extended version enforces non-literal values for these properties, ensuring greater interoperability and alignment with linked data best practices.

In the following sections, we present the components of the Extended Almes Core. Fig. 3.1 in Chapter 3 gives the conceptual model we used as foundation to develop Almes Core as a FAIR metadata schema. This dissertation focuses on the extended form of Almes Core, as it encompasses all the classes and properties included in the simple form. Moreover, the extended version was extensively utilized in the case study described in this research, making it the primary focus of our explanation.

5.2.1 Conceptual Model

The Extended Almes Core conceptual model is shown in Fig. 5.1. It encompasses the description of the metadata templates and the dataset template detailed in Section 4.4. The class Observation now describes

---

[1] `https://almescore.github.io/Almes-Core/`

the data points in the price index datasets. These data points consist of a date (indicating when the price was published), a price (the monetary value), and the currency in which the price is given (e.g., BRL, USD). Each Observation is associated with a Dataset through the `alm:hasObservation` property.

Almes Core is designed as a star schema, allowing multiple Observations to be associated with the same dataset. The `alm:hasObservation` property associates datasets and their recorded observations. The metadata describing the datasets and observations comprises properties from the Resource, Dataset, Agent, and GeoNamesFeature classes.

The `publisher` and `creator` properties describe the entities responsible for publishing the Datasets. To provide more detailed information about these entities beyond just their names, these properties can be instantiated using the `Agent` class, which includes attributes such as name, homepage, and email. Using the `Agent` class is optional in the Simple Almes Core, and mandatory in the Extended Almes Core.

As mentioned before, in the Simple Almes Core, `location` may provide only the name of the location in textual form. In Extended Almes Core, `location` should be instantiated using the `GeoNamesFeature` class, which has two properties: `geoNamesName`, to give the location name as it appears on the GeoNames ontology, and `geoNamesID`, to provide the URI of the concept representing a location within the GeoNames ontology.

The `Resource` and `Dataset` classes should be used together to create the metadata record describing the datasets. In DCAT, the primary distinction between the class Resource and the class Dataset lies in their scope and usage. Resource is a broad class representing any kind of data or metadata resource, including datasets, distributions, services, and other resources published or curated by an agent [130]. In contrast, Dataset is a subclass of Resource that specifically represents structured data collections, which can be broken down into distributions and are managed by data providers. Essentially, while all datasets are resources, not all resources are datasets. The Dataset class is more specialized within the general Resource class in DCAT [130]. In the context of Almes Core, there is no distinction between Resource and Dataset, since their properties are combined in the same metadata record.

The *Product group* property has been defined to group data records into broader product categories, while the *Product name* property specifies the subgroup of the product. In a metadata record about carrots, for example, the Product group would be vegetable, and the product name would be carrot. We are using the Agrotermos thesaurus [132] to represent data for product group and name properties. However, although Agrotermos has a large base of agriculture terminology, some important terms from the trade’s domain are still missing, such as ‘boi gordo’ (Portuguese for fed cattle), which is a very popular term to desig-

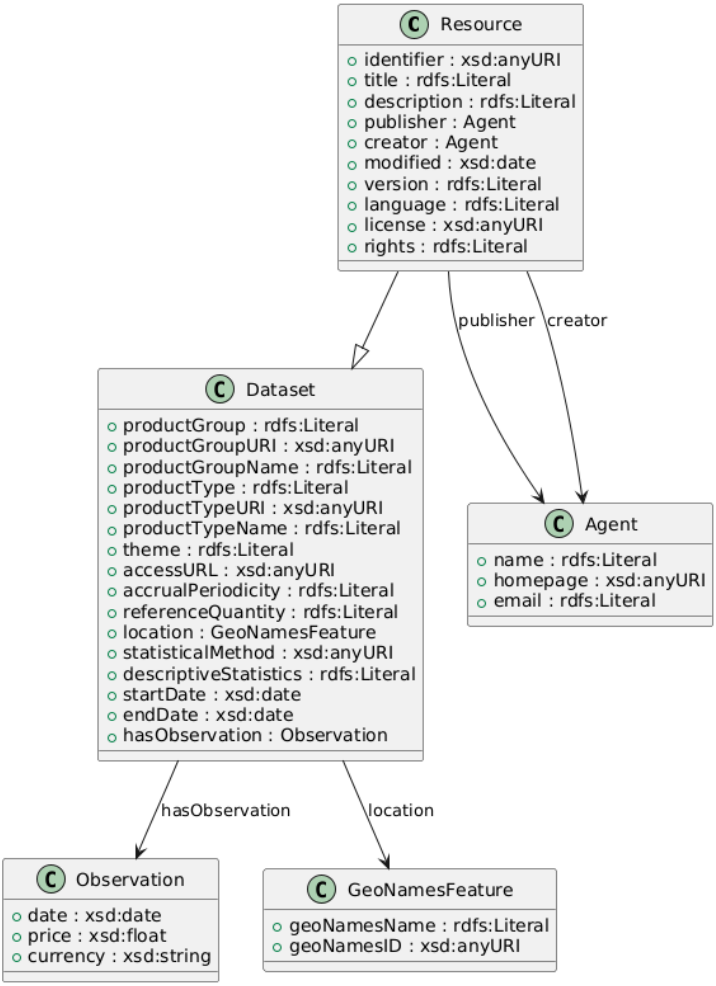


Figure 5.1 Extended Almes Core Conceptual Model [131].

nate a specific kind of cattle in the Brazilian trade market. Therefore, we developed a new ontology, described in subsection , to better represent this domain.

Besides the controlled vocabulary fields, the *verbatim name* property allows the name given in the original dataset to be preserved. The Theme property indicates the kind of trade indicators, which is deeply related to the methodology property since each kind of trade indicator follows a possibly different method.

Publisher and *Creator* were imported from the Dublin Core schema. The Publisher property can be used to refer to the institution that published the Datasets, e.g., CEPEA, Ipea, CONAB. In contrast, the Creator property refers more specifically to the person or department within those institutions responsible for generating the Datasets. The References property was also imported from Dublin Core and gives the link to the address of the dataset on the Web. The License property was another Dublin Core term we incorporated into our metadata schema. It describes the license of use of the dataset given by the data provider.

The metadata schema presented in this chapter has been named Almes Core. The name hints at a broader scope than the metadata schema presented in this chapter, which only focuses on trading data. This is because we intend to extend the metadata schema to cover other sub-disciplines of agriculture and livestock in the upcoming years of the project.

5.2.2 Schema Authority

The Almes Core authority, named the Almes Core Management Group, consists of researchers from institutions in Brazil and the Netherlands, including domain specialists and semantic web experts. Their names and ORCIDs are declared on the schema landing page and in the machine-readable metadata record describing the schema.

The Almes Core Management Group conducted several meetings from 2021 to 2024 to discuss the concepts and the conceptual model of Almes Core. Workshops were organized during the 5th Brazilian Action Plan on Open Government to expose the metadata schema to external evaluation. The group will continue to support the schema's development and growth in the future, as it is expected to be extended to cover other related domains in agriculture.

5.2.3 Landing Page

The Almes Core landing page consolidates all the information necessary to properly use the schema. It is available as a GitHub page[2], which is also redirected by the schema namespace URI[3]. This landing page utilizes the HTTP protocol, which is a fundamental protocol for data exchange on the web, ensuring interoperability and accessibility across different systems and platforms.

The landing page includes a human-readable HTML metadata record describing the schema, as well as a embedded machine-readable metadata record in JSON-LD for automated retrieval purposes (e.g., to be indexed by search engines).

Additionally, the landing page contains a comprehensive list of terms with textual definitions for each term in the schema, including the rules

---

[2] https://almescore.github.io/Almes-Core/

[3] https://w3id.org/AlmesCore#

and relationships between terms. All the namespaces of other schemas reused by Almes Core are also declared on this landing page. This declaration ensures clarity and traceability of the reused vocabularies, promoting semantic consistency and reuse.

The vocabularies reused by Almes Core are listed in Table 5.1. These vocabularies include well-established ontologies and metadata schemas, ensuring that Almes Core adheres to widely accepted practices in the semantic web community. The alm prefix refers to the Almes Core namespace, and dc, sdo, dcat, dct, and foaf refer to other vocabularies from which Almes Core boroughs elements. Agrovoc, Agrotermos and APTO refer to ontologies which their terms are recommended by controlled vocabulary fields within Almes Core. In addition to that, Almes Core has been published on Agroportal[4] for improving findability and reuse.

Table 5.1 Namespaces of Metadata Schemas and Vocabulary Encoding Schemes Reused by Almes Core

Name/Prefix	Namespace
alm	https://w3id.org/AlmesCore#
almrdf	https://w3id.org/AlmesCore/rdf#
dc	http://purl.org/dc/elements/1.1/
sdo	https://schema.org/
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/
foaf	http://xmlns.com/foaf/0.1/
gn	http://www.geonames.org/ontology#
Agrovoc	http://aims.fao.org/aos/agrovoc
Agrotermos	https://sistemas.sede.embrapa.br/agrotermos/
APTO	https://w3id.org/apto

5.2.4 Managing Namespaces

Managing the Almes Core namespace initially proved challenging due to our choice of PURL for handling URIs, as the platform was not functioning reliably. Requests to add new URIs or update target URLs were frequently unprocessed, requiring repeated attempts. Despite multiple efforts, we were unable to implement these changes successfully. This issue was not unique to our experience; several users have reported similar problems, including significant delays and unresolved requests, indicating broader maintenance and reliability concerns with the platform.

---

[4] `https://agroportal.lirmm.fr/ontologies/ALM`

In early 2024, a hacker attack[5] further disrupted the PURL service, exposing vulnerabilities in its infrastructure and rendering it temporarily inaccessible. Given these persistent challenges, we decided to migrate to W3ID [133], a service maintained by the W3C Permanent Identifier Community Group, which offers a more stable and well-supported environment for managing permanent URIs. Consequently, the Almes Core namespace became `https://w3id.org/AlmesCore#`.

To create persistent and reliable URIs for Almes Core using W3ID, we forked the perma-id/w3id.org repository, created a dedicated directory for Almes Core, and configured redirection via an .htaccess file. After synchronizing with the original repository, we submitted a pull request to merge these changes, ensuring that the W3ID URIs correctly pointed to the schema documentation hosted on GitHub.

5.2.5 Metadata Record

To enhance the findability and interoperability of Almes Core, structured metadata was embedded directly into the GitHub Pages site[6] using JSON-LD. This approach ensures that search engines, data aggregators, and semantic web technologies can efficiently index and interpret the metadata.

The metadata record was added as a `<script>` element in the `index.md` file of the GitHub Pages site. This metadata is not visible on the page but is readable by search engines and metadata crawlers.

To confirm the effectiveness of this approach, the site was tested using Google Rich Results Test[7]. These tests successfully detected the embedded metadata[8], verifying that it is correctly formatted and qualifies for enhanced indexing in search engines.

5.2.6 Metadata Elements

The Extended Almes Core includes the elements shown in Table 5.2

Table 5.2 Schema Elements with Domain and Range.

Term Name	Type of Term	Label	Domain	Range
dcat:Resource	Class	Resource	-	-
dct:identifier	Property	Identifier	dcat:Resource	`xsd:anyURI`
dc:title	Property	Title	dcat:Resource	`rdf:Literal`
dct:description	Property	Description	dcat:Resource	`rdf:Literal`

Continued on next page

---

[5] `https://www.reddit.com/r/semanticweb/comments/1galfob/purl_is_down/`

[6] `https://almescore.github.io/Almes-Core/`

[7] `https://search.google.com/test/rich-results`

[8] Refer to the results on `https://search.google.com/test/rich-results/result?id=deilDMsabng6Bz5yzSu4zg`

(Continued from previous page)

Term Name	Type of Term	Label	Domain	Range
dc:publisher	Property	Publisher	dcat:Resource	foaf:Agent
dc:creator	Property	Creator	dcat:Resource	foaf:Agent
dct:modified	Property	Modified	dcat:Resource	`xsd:date`
dct:hasVersion	Property	Version	dcat:Resource	`rdf:Literal`
dct:language	Property	Language	dcat:Resource	`rdf:Literal`
dct:license	Property	License	dcat:Resource	`xsd:anyURI`
dc:rights	Property	Rights	dcat:Resource	`rdf:Literal`
dcat:Dataset	Class	Dataset	-	-
alm:productgroup	Property	Product Group	dcat:Dataset	`rdf:Literal`
alm:productgroupuri	Property	Product Group URI	dcat:Dataset	`xsd:anyURI`
alm:productgroupname	Property	Product Group Name	dcat:Dataset	`rdf:Literal`
alm:producttype	Property	Product Type	dcat:Dataset	`rdf:Literal`
alm:producttypeuri	Property	Product Type URI	dcat:Dataset	`xsd:anyURI`
alm:producttypename	Property	Product Type Name	dcat:Dataset	`rdf:Literal`
alm:theme	Property	Theme	dcat:Dataset	`rdf:Literal`
dcat:accessURL	Property	Access URL	dcat:Dataset	`xsd:anyURI`
dct:accrualPeriodicity	Property	Periodicity	dcat:Dataset	`rdf:Literal`
sdo:referenceQuantity	Property	Reference Quantity	dcat:Dataset	`rdf:Literal`
sdo:location	Property	Location	dcat:Dataset	gn:Feature
alm:statisticalMethod	Property	Statistical Method	dcat:Dataset	`xsd:anyURI`
alm:descriptiveStatistics	Property	Descriptive Statistics	dcat:Dataset	`rdf:Literal`
sdo:startDate	Property	Start Date	dcat:Dataset	`xsd:date`
sdo:endDate	Property	End Date	dcat:Dataset	`xsd:date`
alm:hasObservation	Property	Has Observation	dcat:Dataset	alm:Observation
foaf:Agent	Class	Agent	-	-
foaf:name	Property	Name	foaf:Agent	`rdf:Literal`
foaf:homepage	Property	Homepage	foaf:Agent	`xsd:anyURI`
foaf:mbox	Property	E-mail	foaf:Agent	`rdf:Literal`
gn:Feature	Class	GeoNames Feature	-	-
gn:name	Property	GeoNames Name	gn:Feature	`rdf:Literal`
gn:geonamesID	Property	GeoNames ID	gn:Feature	`xsd:anyURI`
alm:Observation	Class	Observation	-	-
dc:date	Property	Date	alm:Observation	`xsd:date`
sdo:price	Property	Price	alm:Observation	`xsd:float`

Continued on next page

(Continued from previous page)

Term Name	Type of Term	Label	Domain	Range
sdo:currency	Property	Currency	alm:Observation	`xsd:string`

In Almes Core, metadata elements are formally defined as either classes or properties. Each property is associated with a class as its domain. Some properties are defined as subproperties of others, such as alm:productgroup and alm:productgroupname. For the range, each property is assigned a datatype for the expected value.

In addition to domain and range, which define the relationships between elements, other attributes support the semantic and content specification of each element. Each metadata element has a name, based on the recommendations of ISO/IEC 11179-5:2015 [103], as discussed in Chapter 3. Each element also has a label in English and Portuguese for human readability. Each element has its own URI within the Almes Core namespace, and for elements imported from other vocabularies, the original vocabulary's URI is maintained. The definition attribute further complements the semantic specification of the element, providing details about its meaning and scope. Additionally, the comment attribute provides further details regarding the expected contents (values) for the metadata field.

5.2.7 Machine-Actionable Serializations

The following serializations have been deployed:

- JSON-LD[9]
- OWL[10]
- SHACL.ttl [11]
- PlantUML[12]

These serializations are available on GitHub, and a backup on Zenodo [131]. The JSON-LD and OWL serializations encode the full Almes Core schema, including all classes, their properties, and relationships. The SHACL serialization includes only the classes Resource, Dataset, and Agent. It was designed to create metadata records for datasets to be indexed on the FAIR Data Point platform, which indexes metadata

---

[9] `https://raw.githubusercontent.com/AlmesCore/Almes-Core/refs/heads/main/Serializations/alm.json`

[10] `https://raw.githubusercontent.com/AlmesCore/Almes-Core/refs/heads/main/Serializations/alm.xml`

[11] `https://raw.githubusercontent.com/AlmesCore/Almes-Core/refs/heads/main/Serializations/alm-shacl.ttl`

[12] `https://raw.githubusercontent.com/AlmesCore/Almes-Core/refs/heads/main/Serializations/alm.plantuml`

descriptions rather than the data itself. Consequently, this serialization does not include the Data class from the Almes Core schema. The HTML serialization was designed for humans, and the PlantUML to allow exporting a conceptual model diagram for the schema.

5.2.8 Assessing FAIRness

We conducted a FAIR assessment of Almes Core to validate the concepts presented and discussed in chapter 3. For this assessment, we looked for the elements from the ontological conceptual model on the schema's landing page, available at `https://w3id.org/AlmesCore#`. The results of this FAIR assessment are available in details on Zenodo [9].

The FAIR principles were assessed based on the presence or absence of specific FAIR enabling resources, with each resource receiving a score from 0.25 to 1. These FAIR enabling resources were defined based on the conceptual model shown in Fig. 3.1, in Chapter 3, which was derived from the refined FAIR Data principles presented in Table 3.1. The scores for each resource within a principle were then summed to calculate the total score for that principle. Details on the scores of each element are shown in Table 5.3.

Table 5.3 Template for Calculating the FAIRness Level.

Ref. FAIR	FAIR P.	Concept present	Concept absent	Score
RF1	F	Namespace (0.5); version unique identifier (0.5)		1
RF2	F	Landing page (0.5); metadata record for humans (0.25); metadata record for machines (0.25)		1
RF3	F	Indexing in vocabulary catalog (1)		1
RF4	A	URLs use universally accessible protocols (1)		1
RF5	A	Backup (1)		1
RF6	I	Machine-actionable serialization (1)		1
RF7	I	Schema Conceptual model (1)		1
RF8	I	Data Properties (0.5); Range and Domain (0.5)		1
RF9	I	Object Properties (0.5); Range and Domain (0.5)		1
RF10	R	Term name (0.5); Definition (0.5)		1
RF11	R	Open License (1)		1
RF12	R	Documentation of Modifications (1)		1
RF13	I (50%) R (50%)	Schema reuse (1)		1

Each component related to a principle was scored based on its presence (true) or absence (false), based on the method for FAIR assessment of ontologies presented by [91]. If a component is present, it contributes with its full score; if absent, it contributes with zero. For example, for Principle F (Findability):

```
RF1 = Namespace (0.5) (true) + version unique identifier (0.5) (
    ↪ true)
RF1 = 0.5 + 0.5 =1

RF2 = Landing page (0.5) (true) + metadata record for humans (0.25)
    ↪  (true) + metadata record for machines (0.25) (true)
RF2 = 0.5 + 0.25 + 0.25 = 1

RF3 = Indexing in vocabulary catalog (1) (true) = 1
RF3 = 1
```

The total score for each principle is the sum of its components' scores:

```
Total Score for F = RF1 + RF2 + RF3 = 1 + 1 + 1 = 3
```

The fulfillment percentage for each principle (PF%) is calculated by dividing the total score (TS) by the maximum possible score (MPS) for that principle, then multiplying by 100.

$$\mathrm{PF\%} = \left(\frac{\mathrm{TS}}{\mathrm{MPS}}\right) \times 100 = \left(\frac{3}{3}\right) \times 100 = 100\%$$

The overall FAIR compliance score (OFC%) is the average of the fulfillment percentages of all principles. the overall FAIRness score of the Almes Score schema was 100%:

$$\mathrm{OFC\%} = \frac{100\% + 100\% + 100\% + 100\%}{4} = 100\%$$

Overall, the schema performed well in all FAIR aspects. However, the proposed FAIR assessment requires further refinement. Developing an automated tool to streamline the assessment process would be a valuable improvement, as manual evaluation is inherently subjective and may introduce biases. Future work should include additional empirical validation through multiple case studies and user feedback from metadata schema designers to further enhance and validate this conceptual model.

5.3 Mapping the Datasets to the Almes Core Schema

After publishing Almes Core, we mapped column names from the templates described in Section 4.4 to the Extended Almes Core. The mapping of the metadata template is shown in Table 5.4, and the mapping of the data template in Tables 5.5 and 5.6. For the Datasets in which the date was a range, the columns were mapped as separated fields as shown in Table 5.6.

Table 5.4 Mapping of Elements from the Metadata Template to Extended Almes Core.

Metadata template	Almes Core
id	dct:identifier
grupoProdutoNome	alm:productGroupName
grupoProdutoURL	alm:productGroupURI
produtoNome	alm:productTypeName
produtoURL	alm:productTypeURI
titulo	dc:title
Descricao	dct:description
editor	dc:publisher
criador	dc:creator
referencia	dcat:accessURL
estatistica descritiva	alm:descriptiveStatistics
periodicidade	dct:accrualPeriodicity
tema	alm:theme
unidade	sdo:referenceQuantity
localizacao	gn:name
localizacaoID	gn:geonamesID
periodoReferenciaInicio	sdo:startDate
periodoReferenciaFim	sdo:endDate
licenca	dct:license
direitos	dc:rights
metodologia	alm:statisticalMethod

Table 5.5 Mapping of Elements from the Data Template to Almes Core, with a Single Date as Reference for the Period.

Data template (Observation)	Almes Core
metadata_id	dct:identifier
date	dc:date
price	sdo:price
currency	sdo:currency

Table 5.6 Mapping of Elements from the Data Template to Almes Core, with a Range of Dates as Reference for the Period.

Data template	Almes Core
metadata_id	dct:identifier
startDate	sdo:startDate
endDate	sdo:endDate
price	sdo:price
currency	sdo:currency

After mapping the column names, we began the process of aligning the dataset values with the data types specified by Almes Core. Some metadata fields from the metadata template have been populated with terms from controlled vocabularies instead of literal values as in the original datasets. Starting with `alm:productGroup`, the two subproperties `alm:productGroupURI` and `alm:productGroupName` from the Extended Almes Core were used to include both the name and the URI of the product group. The same approach was applied to `alm:productType`.

During our search for linked data controlled vocabularies that could comprehensively represent all product types and groups included in the datasets, we found that no existing controlled vocabulary could cover all these concepts. Given this limitation, we developed a new ontology named APTO, described in Chapter 6, to serve as the vocabulary encoding schema for these two metadata fields.

The fields `alm:descriptiveStatistics` and `alm:theme` were populated with literal values from the original datasets. Almes Core did not recommend any specific controlled vocabulary for these fields. This enhancement is expected to be included in an upcoming version of the schema.

The values for the `dct:accrualPeriodicity` field were mapped to terms from the Collection Description Frequency Vocabulary (DCMI-COLLFREQ) [134], as recommended by Almes Core.

Location values were mapped to GeoOntology concepts through the fields gn:name and gn:geonamesID,providing a systematic approach to reference geographical information.

Additionally, the `dct:license` field includes URLs from the Creative Commons Initiative as values, ensuring that the licensing information is both accessible and standardized. This practice aligns with the broader goals of open data and FAIR principles, promoting transparency and usability of the data. In Chapter 6, we describe the development of APTO. In Chapter 9, we detail C4AI-KGAP, in which all components of the framework are brought together.

CHAPTER 6

The Agricultural Product Types Ontology

Chapter Overview

This chapter introduces APTO, an ontology designed to classify agricultural products while aligning with FAIR data principles. The chapter begins by discussing the limitations of existing classification systems for agricultural product types (Section 6.1), highlighting their inconsistencies and lack of interoperability. We then present a comparative analysis of agricultural products classification systems, including Mercosur Common Nomenclature (NCM) and the United Nations Statistics Division (UNSD) classifications. In Section 6.5 we present APTO structure, which consists of two main modules: (1) the Product Types module, which classifies products based on their origin (animal, plant) and processing stage (raw, processed, by-product), and (2) the Organism module, which links products to their biological taxonomy, integrating with the GBIF Backbone Taxonomy. In Sections 6.3, 6.4, and 6.5 we presented the ontology development process, which followed the SABiO methodology. Finally, we describe APTO's implementation and integration (Section 6.6).

6.1 Classification Systems on Agricultural Product Types

Every country in the world adopts its own schemas to classify the products resulting from agricultural economic activities. In Brazil, this is done by the 'Nomenclatura Comum do MERCOSUL (NCM)', which standardizes the names of agricultural products not only for Brazil but for all South-American countries that participate in the MERCOSUL commercial alliance. This classification is published as a series of tables on the Brazilian Ministry of Economic Affairs website [135].

The need for a unified ontology for classifying agricultural products is further confirmed by various existing classification systems worldwide, which, while effective for more basic uses (e.g., consulting product names in a list), often fail to meet the FAIR Data principles and Semantic Web standards, since the format they are published hinders findability, accessibility, Interoperability, and reuse. This section reviews several classification systems to highlight the diversity and the need for a ontology.

The United Nations Statistics Division (UNSD) provides a comprehensive classification system for economic statistics, including agricultural products. The UNSD classifications are designed to standardize economic data across countries, facilitating comparability and integration into global economic statistics. This framework includes detailed categorizations for various agricultural products, thereby supporting international trade and economic analysis [136].

In Europe, agricultural products are classified to facilitate pesticide residue assessment. This system covers various aspects such as plant metabolism, crop sampling, residue trials, feed, and validation of analytical methods for food safety evaluation [137, 138]. Similarly, the United States Department of Agriculture (USDA) classifies field crops based on plant characteristics, purpose of production and consumption, plant longevity, environmental needs, and agricultural operations [139]. These classifications aid in efficient production processes and promote food security worldwide.

Another classification approach focuses on the commercial value and quality standards of agricultural products. For instance, in Brazil, products are grouped into types based on the occurrence of defects in accordance with established standards and legislation [140].

In another example, the liberalized marketing environment in Bangladesh highlights another aspect of agricultural product classification for trading. This system benefits both producers and consumers by allowing appropriate market access for farmers and efficient price information flow from traders to consumers. This approach is particularly effective for trading high-value agricultural products like vegetables, fish, poultry, and milk [141].

The effects of regulatory heterogeneity and the relative rigor of agri-food standards on agricultural products trading are also significant. For instance, the impact of regulations associated with Maximum Residue Limits (MRLs) of pesticides can influence compliance costs, trade flows, and potentially allow for higher unit prices if exporters meet these regulatory requirements [142].

Despite the effectiveness of these systems, most classifications are published as lists in PDF format or spreadsheets, which are not conducive to reusability in the Semantic Web. For instance, the Brazilian NCM classification is published as tables in PDF and CSV files. In this classification, product types are categorized in sections and subsections (chapters). Each entry is a long list of names, rather than a single term representing a single concept. This approach can be confusing and meaningless from an ontology modeling perspective, although it is common practice for classification systems. Other classification systems, such as the Universal Decimal Classification (UDC) and Dewey Decimal Classification (DDC), also adopt this style, referring to each entry as a subject heading instead of a term or concept, as in ontology modeling. An example of the first section of the NCM classification is shown in Listing 6.1.

Listing 6.1 Snippet from the NCM Classification

```
Section I -LIVE ANIMALS AND PRODUCTS OF ANIMAL ORIGIN
   Chapter 01: Live animals.
   Chapter 02: Meat and edible offal.
   Chapter 03: Fish and crustaceans, mollusks, and other aquatic
       ↪ invertebrates.
   Chapter 04: Milk and dairy products; birds' eggs; natural honey;
       ↪ edible products of animal origin,
   not specified or included in other Chapters.
   Chapter 05: Other products of animal origin, not specified or
       ↪ included in other Chapters.
```

Listing 6.1 shows that products of different natures, such as "milk and dairy products" and "birds" eggs", are grouped together in the same category. From an ontological perspective, these types of products should be separated and represented by different terms, since they represent distinct concepts. Another issue with this classification is that not all products of animal origin are grouped in the same category, and the same happens with products of plant origin. For instance, "fat and animal oil" is a subclass of "Section III - Fats and oils of animal, vegetable, or microbial origin and products of their dissociation; processed edible fats; waxes of animal or vegetable origin." Section III groups products based on their type (oil) despite their different origins (animal, plant, or microorganisms). Thus, while the first two categories of the system focus on the biological aspect of the products producers (animals and plants), other sections group the product types according to the type of processing.

In an ontology, this would not be an issue, as ontologies allow polyhierarchy, enabling multiple categorizations for a single concept. However, the current structure of this classification system does not allow any semantic relationships between concepts to be traced. This limitation demonstrates the need for a more flexible and semantically rich classification system, to enhance data interoperability and reusability in the Semantic Web.

Similarly, the UN Classifications on economic statistics [136], even though available in a more structured JSON format, also publish their classification system with long subject headings instead of concepts and relationships, to show these different aspects of the product types, as seen in Listing 6.2.

Listing 6.2 Snippet from the UN Classifications on Economic Statistics

```
{
      "id": "01",
      "text": "01 - Animals; live",
      "parent": "TOTAL",
      "isLeaf": "0",
      "aggrlevel": 2,
      "standardUnitAbbr": "n/a"
},
{
      "id": "0101",
      "text": "0101 - Horses, asses, mules and hinnies; live",
      "parent": "01",
      "isLeaf": "0",
      "aggrlevel": 4,
      "standardUnitAbbr": "n/a"
},
{
      "id": "010121",
      "text": "010121 - Horses; live, pure-bred breeding animals",
      "parent": "0101",
      "isLeaf": "1",
      "aggrlevel": 6,
      "standardUnitAbbr": "u"
},
{
      "id": "010129",
      "text": "010129 - Horses; live, other than pure-bred breeding animals",
      "parent": "0101",
      "isLeaf": "1",
      "aggrlevel": 6,
      "standardUnitAbbr": "u"
}
```

This classification system is more structured than the Brazilian NCM, as shown in Listing 6.1. Each entry has an ID, which is used to determine the hierarchical relationships using the `parent` property. However, it still relies on long subject headings rather than formal concepts and relationships, limiting its interoperability.

The limitations of these traditional classification systems id that they lack the ability to represent complex relationships between concepts and do not adhere to the principles of Semantic Web. For example, the NCM and UN classifications group diverse products under broad categories, which can obscure important distinctions and relationships.

An ontology-based approach addresses these issues by using a language with formal semantics such as, e.g., OWL. This allows for more precise and flexible data representation. For instance, an ontology can separate milk and birds' eggs into distinct classes and define their relationships to other entities, such as the organisms that produce them, as shown in Listing 6.3.

Listing 6.3 Excerpt of an OWL Representation of NCM Classification

```
<owl:Class rdf:about="#LiveAnimal"/>
<owl:Class rdf:about="#AnimalProduct"/>
<owl:Class rdf:about="#Milk">
   <rdfs:subClassOf rdf:resource="#AnimalProduct"/>
</owl:Class>
<owl:Class rdf:about="#BirdEgg">
   <rdfs:subClassOf rdf:resource="#AnimalProduct"/>
</owl:Class>
<owl:Class rdf:about="#Horse">
   <rdfs:subClassOf rdf:resource="#LiveAnimal"/>
</owl:Class>
<owl:ObjectProperty rdf:about="#produces">
   <rdfs:domain rdf:resource="#Animal"/>
   <rdfs:range rdf:resource="#AnimalProduct"/>
</owl:ObjectProperty>
<owl:NamedIndividual rdf:about="#DairyCow">
   <rdf:type rdf:resource="#Animal"/>
   <produces rdf:resource="#Milk"/>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="#Hen">
   <rdf:type rdf:resource="#Animal"/>
   <produces rdf:resource="#BirdEgg"/>
</owl:NamedIndividual>
```

6.2 Ontologies for Agricultural Product Types

A quick search for the term “product” on Agroportal [143], which is the most extensive ontology repository for agriculture, yielded 13 ontologies that included this term as of July 31, 2024. Among these, only five ontologies included a categorization of product types:

1. The Livestock Product Trait Ontology (LPT) is defined as a controlled vocabulary for describing traits (measurable or observable characteristics) pertaining to products produced by or obtained from the body of an agricultural animal or bird maintained for use and profit [144]. This ontology focuses on traits such as morphological characteristics, composition, shape, and size rather than on product types.

2. Crop Planning and Production Process Ontology (C3PO) [145] was developed to assist vegetable crop farmers in planning and monitoring their agricultural activities. It captures the complexity of diversified crop production through eight modules that represent farming practices and constraints. C3PO encodes agricultural processes, farm plot organization, and farming tasks, from soil preparation to harvest. Serving as the backbone of a knowledge graph, C3PO integrates data from various semantic resources, enhancing decision support systems for farmers and promoting sustainable

farming practices. Notably, this ontology includes the representation of plants and their botanical families, which is similar to the approach we adopted to develop APTO.

3. The Meat Thesaurus (MEAT-T) is a controlled vocabulary on meat and meat products [146]. MEAT-T encompasses a wide range of terms describing different types of meat, meat cuts, and related processing techniques. It facilitates consistent communication and data interoperability across various domains, including research, industry, and regulatory sectors. By providing a structured set of terms, MEAT-T enhances the ability to share and compare data on meat products, contributing to improved traceability, quality control, and market analysis.
4. The Compositional Dietary Nutrition Ontology (CDNO) represents nutritional attributes of food-related material entities, focusing on quantifying chemical nutritional components from various stages of food production, including crops, livestock, and fisheries, through processing and supply chains [147]. The top class called "dietary nutritional component" organizes terms in a hierarchy reflecting commonly used classifications of chemical food composition, while terms within the "nutritional component concentration" class represent quantification data. CDNO facilitates the description, communication, and exchange of nutritional composition data, supporting applications in research, food composition tables, labeling, supply chain quality control, and biodiversity management. It has been designed to be used in conjunction with the class "Food product by organism" from the Food Ontology (FoodOn), which relates products to the NCBITaxon [147].
5. The Ontology for Food Processing Experiment (OFPE) is a generic ontology designed to represent products and activities during food transformation processes [148]. It is structured around four main concepts: Product, Operation, Attribute, and Observation. OFPE provides a framework for classifying products according to their processing methods but does not aim to be a comprehensive classification system for product types. Instead, it focuses on categorizing the processes and observations related to food transformation, facilitating better understanding and communication in food processing research and industrial applications.

Three other controlled vocabularies not retrieved in our search on Agroportal offer comprehensive classifications for product types: FoodOn [149], Agrovoc, and Agrotermos [132]. Among these, FoodOn is notable for its detailed and complex categorization. However, FoodOn focuses exclusively on food products and does not cover non-food items such as natural rubber and wood, making it incomplete for this case study. Consequently, we based APTO on the structure of product types found in Agrovoc and Agrotermos, which provide a more comprehensive classification of agricultural product types. In addition, Agrotermos is the only controlled vocabulary specifically representing concepts of the Brazilian agriculture. As of 2022, Agrotermos included 245,000 terms, 41,337 of which were incorporated from Agrovoc [150].Our decision to work with Agrotermos also aimed to enhance the Brazilian vocabulary, as APTO was developed to be integrated into Agrotermos.

Agrotermos encompasses most of the 218 product types identified across the CEPEA, Ipea, and CONAB databases, excluding only 79 types. Since Agrotermos is the official Brazilian Portuguese terminology provider for Agrovoc, we also incorporated classes from Agrovoc, which serves a global audience. While we did not directly reuse classes from FoodOn, we drew inspiration from its modeling approach. For instance, APTO relates product types to their biological aspects through taxonomic classification provided by the Global Biodiversity Information Facility (GBIF), similar to how FoodOn uses the NCBI taxonomy [147].

Despite the inclusion of many product types from the Brazilian price index databases by these ontologies, none covered all 218 product types identified across the CEPEA, Ipea, and CONAB databases. Many of these products are derived from species endemic to Brazil and are not cultivated or raised elsewhere. This is why we developed APTO as a new ontology, reusing existing ontologies, mainly Agrovoc and Agrotermos.

6.3 Development Process

We developed APTO in accordance with the SABiO's Development Methodology [73], which is shown in Fig. 2.5, Chapter 2.

The SABiO methodology was designed to guide the development of ontologies, supporting both reference and operational ontologies [73]. The methodology is structured into five main phases, each with associated processes and subprocesses. These phases and their key components are the following:

1. **Purpose Identification and Requirements Elicitation**
    - Purpose and Intended Uses Identification
    - Requirements Elicitation
    - Competency Questions Identification
    - Ontology Modularization
2. **Ontology Capture and Formalization**
    - Conceptual Modeling
    - Dictionary of Terms Definition
    - Informal Axioms Definition
    - Formal Axioms Definition

3. **Design**
    - Technical Non-Functional Requirements Elicitation
    - Implementation Environment Definition
    - Architectural Design
    - Detailed Design
4. **Implementation**
    - Implementation in the Chosen Operational Language
5. **Testing**
    - Sub-ontology Testing
    - Integration Testing
    - Full Ontology Testing

In addition, the following supporting processes are defined to perform the five phases:

1. Knowledge Acquisition
2. Reuse
3. Documentation
4. Configuration Management
5. Evaluation

The development process is iterative and can follow various life cycle models such as Incremental and Spiral, depending on the project needs [73]. In the sequel, we describe the APTO development in accordance with the SABiO methodology.

6.4 Purpose Identification and Requirements Elicitation

As for the ontology's purpose and intended use, APTO has been designed to serve as a classification system for agricultural price index databases. We categorized the ontology requirement in two groups, following SABiO's recommendations: functional and non-functional requirements [73]. The functional requirements were articulated as two competence questions:

Competence Questions

1. What is the taxonomic classification (species, genus, or family) of the organism from which a given agricultural product originates?
2. Is this product a raw product, a processed product or a by-product?

These two competence questions guided the definition of the ontology's two main modules, Organism and Product Type.

The non-functional requirements are:

- Usability and understandability of the Product Type module to be evaluated using traditional User Experience (UX) measurement techniques.
- The ontology should be the most FAIR as possible, based on requirements for FAIR ontologies elucidated by [151]).
- Different distribution formats are provided, including a normative document in HTML and an OWL serialization. Both formats are synchronized, ensuring that any modification in one is automatically reflected in the other.

6.5 Ontology Capture and Formalization

The main concepts to be represented in APTO were first extracted as a list of product names from CEPEA, Ipea, and CONAB datasets. Then, we mapped all the concepts to a single list, to eliminate synonyms. The final list had 218 product types. After that, we created a first version of the taxonomy for the Product Type module. This taxonomy was based on the Agrovoc/Agrotermos taxonomy of product types. Each product type was also related to the organism that originates that product. To organize the concepts in a logical structure that allowed discussing with specialists, we developed a OntoUML conceptual model before starting the ontology design phase, as suggested by [73].

The ontology is divided into two main modules: `Organism` and `Product Type`.

6.5.1 Organism Module

The `Organism` module shown in Fig. 6.1 provides the foundational biological classification to the entities from which the `Product Types` originate. This module includes the `Organism` class, which is further divided into two main sub-classes:

- `Animalia`: Represents all animal organisms. This class includes any organism that belongs to the animal kingdom and serves as the source for `Animal Products`.
- `Plantae`: Represents all plant organisms. This class includes any organism that belongs to the plant kingdom and serves as the source for `Plant Products`.

The `Animalia` and `Plantae` classes are modeled as disjoint and incomplete, allowing future extensions to include other groups, such as fungi, which may also produce distinct products (e.g., `Fungi Product`).

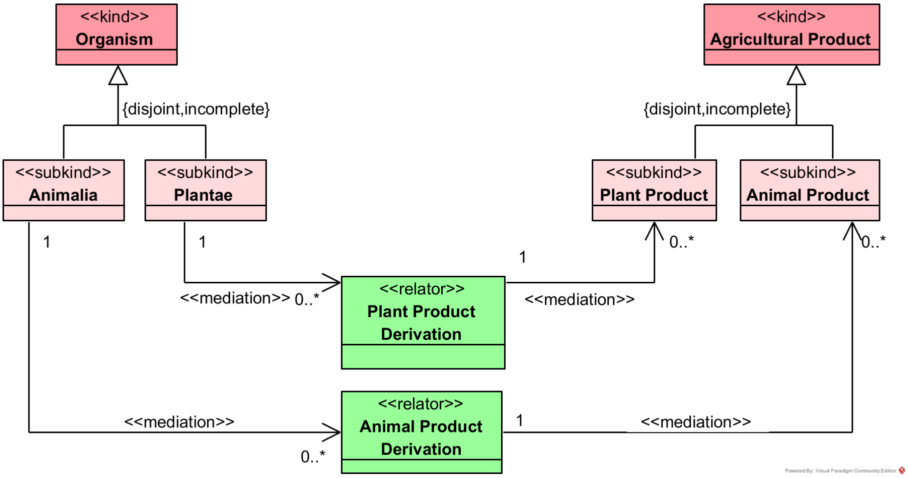


Figure 6.1 Organism Module in OntoUML.

6.5.2 Product Type Module

The `Product Type` class, as shown in Fig. 6.2, is an abstract class representing various categories of products that can be derived from organisms. It is not directly instantiated, but rather serves as a general concept from which more specific product categories are derived. The `Product Type` class has two generalization sets:

- The first generalization set subdivides Product Type into Plant Product and Animal Product, which specify the origin of the product.
  - **Plant Product**: Represents any product derived from a plant. Examples include fruits, vegetables, wood, etc.
  - **Animal Product**: Represents any product derived from an animal. Examples include meat, milk, leather, etc.
- The second generalization set divides Product Type into Raw Product, Processed Product, and By-product, each representing the state or category of the product:
  - **Raw Product**: Represents products in their natural, unprocessed state. These are directly obtained from plants or animals without undergoing any manufacturing process.
  - **Processed Product**: Represents products that have undergone some form of manufacturing or processing, transforming the raw material into a different form or enhancing its value.
  - **By-product**: Represents the residual or secondary products generated during the production or processing of another primary product.

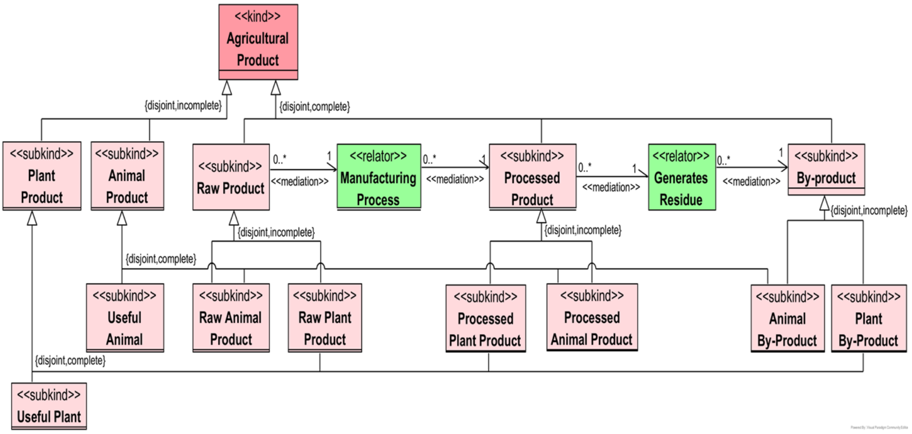

<<kind>> Agricultural Product
{disjoint,incomplete}
{disjoint,complete}
<<subkind>> Plant Product
<<subkind>> Animal Product
<<subkind>> Raw Product
0..*
1
<<mediation>>
<<relator>> Manufacturing Process
0..*
1
<<mediation>>
<<subkind>> Processed Product
0..*
1
<<mediation>>
<<relator>> Generates Residue
0..*
1
<<mediation>>
<<subkind>> By-product
{disjoint,incomplete}
{disjoint,incomplete}
{disjoint,incomplete}
{disjoint,complete}
<<subkind>> Useful Animal
<<subkind>> Raw Animal Product
<<subkind>> Raw Plant Product
<<subkind>> Processed Plant Product
<<subkind>> Processed Animal Product
<<subkind>> Animal By-Product
<<subkind>> Plant By-Product
{disjoint,complete}
<<subkind>> Useful Plant

Figure 6.2 Product Type Module in OntoUML.

At a higher level in the ontology, there are significant relationships between the different types of products:

- A Processed Product is obtained from the Manufacturing Process of a Raw Product. During this process, raw materials are transformed into processed goods, adding value or changing their form to make them more suitable for consumption or use.
- The production of a Processed Product often generates residues or secondary outputs, categorized as By-products. The generation of By-products is an inherent part of the production process.

Animal and Plant Products

The Animal Product and Plant Product classes are further subdivided into more specific sub-classes, combining both the product origin and the type of processing:

- **Animal Product**:
  - **Useful Animal**: Represents living animals that are directly useful to humans, such as livestock (cattle, chicken, etc).
  - **Raw Animal Product**: Represents unprocessed products obtained from animals, such as raw meat, eggs, or wool.
  - **Processed Animal Product**: Represents animal-derived products that have undergone processing, altering their physical or chemical properties, such as cheese, leather, or sausages.
  - **Animal By-product**: Represents secondary products resulting from the processing of animal products, such as whey or gelatin.

- **Plant Product**:
  - **Useful Plant**: Represents whole plants or plant parts that are directly useful to humans without requiring further processing, such as rubber plants or wood trees.
  - **Raw Plant Product**: Represents unprocessed products obtained from plants, such as fresh fruits, vegetables, or raw timber.
  - **Processed Plant Product**: Represents plant-derived products that have undergone processing, altering their physical or chemical properties, such as flour, paper, or refined oil.
  - **Plant By-product**: Represents secondary products resulting from the processing of plant products, such as sawdust or bran.

6.5.3 Relationships Between Product Types and Organisms

The relationships between the Product Type module and the Organism module are as follows:

- A Plant Product is derived from a Plantae entity. This relationship reflects that any product categorized under Plant Product originates from an organism classified under the Plantae class.
- An Animal Product is derived from an Animalia entity. This relationship indicates that any product categorized under Animal Product originates from an organism classified under the Animalia class.

6.5.4 From OntoUML to OWL

While the OntoUML ontology presented in Figs. 6.1 and 6.2 provides a high-level model for representing relationships between organisms and product types, its primary advantage lies in its visual representation, which facilitated discussions with domain specialists. However, it has limitations in capturing more complex relationships within the domain. Therefore, this conceptual model served as inspiration for developing a more comprehensive OWL ontology.

6.6 Ontology Design and Implementation in OWL

APTO contains labels and comments in both Portuguese and English. To design the OWL version of APTO, we utilized Protégé Desktop v.5.6.4 [152]. The majority of classes in the `:Product Type` module were directly imported from Agrovoc and Agrotermos. Initially, we imported only the concept URIs, labels, and hierarchical relationships, with associative relationships being defined later. The `:Organism` module includes subclasses that represent the taxonomy of species associated with the production of specific product types, importing taxonomic information from the GBIF Backbone Taxonomy via API. Other reused ontologies and schemas are listed in the APTO normative namespaces, as shown in Table 6.1, and are also available on the ontology's landing page in [153], as well as in the Data and Code Availability Section [13].

Table 6.1 APTO Namespace Prefixes and IRIs

Prefix	Namespace IRI
apto	`https://w3id.org/APTO#`
ro	`http://purl.obolibrary.org/obo/RO_`
bco	`http://purl.obolibrary.org/obo/BCO_`
owl	`http://www.w3.org/2002/07/owl#`
rdf	`http://www.w3.org/1999/02/22-rdf-syntax-ns#`
sdo	`https://schema.org/`
xml	`http://www.w3.org/XML/1998/namespace`
xsd	`http://www.w3.org/2001/XMLSchema#`
rdfs	`http://www.w3.org/2000/01/rdf-schema#`
skos	`http://www.w3.org/2004/02/skos/core#`

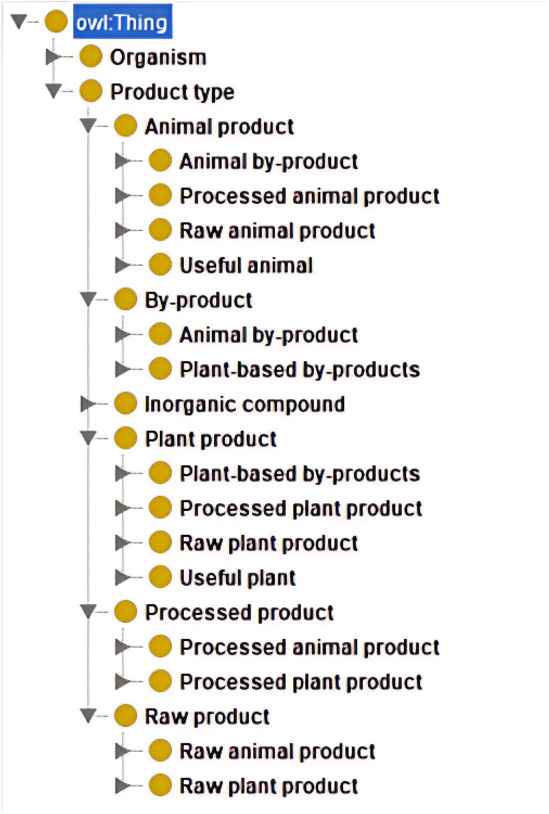
APTO does not include any data properties, only object properties, as it is not intended for direct instantiation with individuals. Default annotation properties from OWL and RDF are utilized, and the SKOS vocabulary is imported to indicate alternative labels for the classes.

6.6.1 :Product_type Module

The :Product_type module is subdivided into two facets: the origin of the product (:Inorganic_compound, :Animal_product, or :Plant_product), and the type of processing (:Raw_product, :Processed_product, :By-product). The classes in the origin facet also subdivide according to the type of processing. This subdivisions are shown in Fig. 6.3.

APTO defines five object properties to indicate other types of relationships between product types:

- :has_ingredient and :is_ingredient_of (inverse properties): Indicate that a product type is used in the manufacturing of a processed product type (e.g., :Cow_cheese :has_ingredient :Cow_milk).
- ro:derives_from and ro:derives_in (inverse properties, imported from the Relations Ontology (RO)): Indicate that a product type originates from another product type (e.g., :Cocoa_almond ro:derives_from :Cocoa).
- :residue_of indicates that a by-product is the residue of manufacturing another product (e.g., :Whey :residue_of :Cheese).



- owl:Thing
  - Organism
  - Product type
    - Animal product
      - Animal by-product
      - Processed animal product
      - Raw animal product
      - Useful animal
    - By-product
      - Animal by-product
      - Plant-based by-products
    - Inorganic compound
    - Plant product
      - Plant-based by-products
      - Processed plant product
      - Raw plant product
      - Useful plant
    - Processed product
      - Processed animal product
      - Processed plant product
    - Raw product
      - Raw animal product
      - Raw plant product

Figure 6.3 APTO Product Type Module in OWL.

Key Differences Between the OntoUML and the OWL Ontologies

The OntoUML ontology presented in Figs. 6.1 and 6.2 is not able to express cases where a processed product could be derived from a combination of both a raw (or primary) product and a by-product. Specifically, the model could not adequately represent scenarios where a product is composed of multiple ingredients with different origins, such as in the case of a Dairy Drink (bebida láctea, in Portuguese), which is produced using both Cow Milk (a raw primary product) and Whey (a by-product) [154].

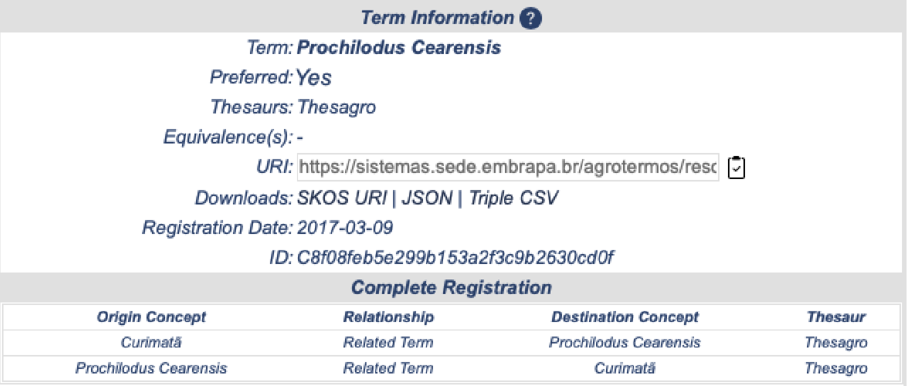
To overcome this limitation, the object properties :has_ingredient and :is_ingredient_of were introduced in the OWL ontology. These properties allow for the explicit representation of products that are composed of multiple ingredients, each originating from different stages of the production process. To model the composition of a Dairy Drink, the following expression in OWL was used:

```
:Dairy_drink has_ingredient some (:Cow_milk and :Whey)
```

This expression specifies that a `Dairy Drink` is derived from the combination of `Cow Milk` and `Whey`, addressing the previously identified modeling limitation. By using these object properties, the model now accurately captures the complexity of products that incorporate multiple types of ingredients.

6.6.2 `:Organism` Module

We chose to import the taxonomic classification for the :Organism module from GBIF. Our review of species names in Agrotermos and Agrovoc revealed the use of outdated taxonomic names, likely due to manual curation that struggle to keep pace with the dynamic nature of taxonomic classification. A common update in scientific taxonomy is synonymization, where different names are found to refer to the same species, and one name (usually the older) is accepted as the correct term [155]. For example, in Agrotermos, we identified that *Prochilodus cearensis* Steindachner, 1911 (Fig. 6.4) is listed as a preferred term, although it is a synonym of *Prochilodus brevis* Steindachner, 1875, according to the GBIF Backbone Taxonomy [156]. Additionally, another problem in the term shown in Fig 6.4 is the format of species names in Agrotermos: it incorrectly capitalizes the specific epithet, which violates the international codes of nomenclature [157, 158].



Term Information

Term: **Prochilodus Cearensis**
Preferred: Yes
Thesaurs: Thesagro
Equivalence(s): -
URI: https://sistemas.sede.embrapa.br/agrotermos/resc
Downloads: SKOS URI | JSON | Triple CSV
Registration Date: 2017-03-09
ID: C8f08feb5e299b153a2f3c9b2630cd0f

Complete Registration

Origin Concept	Relationship	Destination Concept	Thesaur
Curimatã	Related Term	Prochilodus Cearensis	Thesagro
Prochilodus Cearensis	Related Term	Curimatã	Thesagro

Figure 6.4 Synonym Indicated as Preferred Term in Agrotermos [159].

The Organism module in APTO follows a seven-level classification system, covering kingdom, phylum, class, order, family, genus, and species. To streamline the process of modeling the taxonomy in OWL, we experimented with a LLM to automate the inclusion of taxonomic data in APTO, as detailed in Chapter 7.

CHAPTER 7

Ontology Design Supported by an LLM

Chapter Overview

Managing scientific names in ontologies that represent species taxonomies is challenging due to the ever-evolving nature of these taxonomies. Manually maintaining these names becomes increasingly difficult when dealing with thousands of scientific names. To address this issue, this chapter investigates the use of ChatGPT-4 to automate the development of the `Organism` module in APTO for species classification. Our methodology involved leveraging ChatGPT-4 to extract data from the GBIF Backbone API and generate OWL files for further integration in APTO. Two alternative approaches were explored: (1) issuing a series of prompts for ChatGPT-4 to execute tasks via the BrowserOP plugin and (2) directing ChatGPT-4 to design a Python algorithm to perform analogous tasks. Both approaches rely on a prompting method where we provide instructions, context, input data, and an output indicator. The first approach showed scalability limitations, while the second approach used the Python algorithm to overcome these challenges, but it struggled with typographical errors in data handling. This study highlights the potential of LLMs like ChatGPT-4 to streamline the management of species names in ontologies. Despite certain limitations, these tools offer promising advancements in automating taxonomy-related tasks and improving the efficiency of ontology development.

Publication note: This chapter is based on the publication of Soares et al. [160]: *Exploring a Large Language Model for Transforming Taxonomic Data into OWL: Lessons Learned and Implications for Ontology Development.*

7.1 Introduction

Ontologies capture a well-defined and well-expressed shared conceptualization of a domain [161]. They essentially specify the kinds of entities relevant to a particular domain and the relationships among such entities, thereby constructing a large semantic structure that represents the key aspects of domain knowledge [162]. Ontologies also play a crucial role in facilitating the integration of heterogeneous data sources, enabling interoperability among information systems [163].

Ontology development is a comprehensive process involving the conceptualization step, during which domain experts and ontologists work together to specify significant concepts and relationships peculiar to a particular domain. Such collaboration guarantees that the established ontology sufficiently addresses the intricacies of the domain while being responsive to new knowledge and uses [164, 165]. Ontologies facilitate the structured representation of knowledge, enhancing interoperability and reusability in modeling real-world concepts [166].

Ontologies in the agricultural domain often incorporate taxonomic classifications to represent and organize knowledge about organisms with agricultural interest. Biologists define taxonomy as “the theory and practice of identifying, describing, naming, and classifying organisms” [167]. A taxonomy can be considered an ontology, as it formalizes hierarchical relationships and categories among organisms. However, the science of taxonomy faces numerous challenges from the dynamic nature of scientific discovery, as scientists continuously find new species and revise earlier established classifications [168]. The current debate over competing classification systems, such as the Linnaean hierarchy versus phylogenetic taxonomy, further compounds this situation by not reaching a consensus [169]. Additionally, maintaining consistency in naming conventions across languages and disciplines, managing synonyms (multiple names for the same organism), and resolving homonyms (the same name referring to different organisms) are also issues that have to be handled [170].

Despite these challenges, taxonomic classifications remain a valuable resource for structuring agricultural ontologies, aiding in the standardized classification of organisms of agricultural interest. Many approaches exist for integrating taxonomic data into these ontologies, ranging from manual to semi-automated methods. In manually curated vocabularies such as AGROVOC and Agrotermos, curators enter taxonomic names by completing forms to add new concepts, ensuring alignment with authoritative taxonomic databases [171]. In contrast, semi-automated methods, such as those employed by FoodOn, leverage computational tools to streamline taxonomy integration. For example, FoodOn applies a ROBOT template-driven approach to refine seafood nomenclature, facilitating the alignment with NCBI GenBank names and NCBITaxon [172].

Building on existing taxonomy integration approaches, this study evaluates the potential of LLMs to convert taxonomic data into OWL format for integration into an agricultural ontology. Specifically, it leverages ChatGPT-4 to extract taxonomic data from the GBIF Backbone

Taxonomy API, which is then converted into OWL and integrated into the ontology. Two methods are explored: direct API querying via ChatGPT and a Python-based approach using LLM-generated code. The study assesses these methods in terms of scalability, reliability, and effectiveness in incorporating taxonomic data into a domain-specific ontology. In particular, it examines the feasibility of a GPT-driven approach for enhancing APTO by improving taxonomic accuracy and addressing ambiguities in agricultural product names, which often vary across languages, regions, and cultural contexts.

7.2 Related Work

The application of LLMs in KG and ontology development has gained attention due to their ability to process and generate conceptual schemas from natural language text sources [173]. Recent studies have demonstrated that LLMs, such as ChatGPT, can significantly support tasks in ontology engineering, including concept extraction, class hierarchy generation, and semantic relationship identification [173, 174].

A common use of LLMs in ontology development is automating concept extraction and populating ontologies with domain knowledge. For example, Aggarwal et al. [175] experimented with using LLMs in academic ontology development in engineering fields, with promising results in extracting and structuring domain knowledge. Zaitoun et al. [176] also employed LLMs to enrich biomedical ontologies, demonstrating that they are able to identify missing concepts and relations. Trajanoska et al. [177] and Reis et al. [178] demonstrated how LLMs can be utilized for enhancing KG construction from unstructured text.

LLMs also facilitate semantic relationship discovery and ontology alignment. Amini et al. [179] investigated LLMs for complex ontology alignment, using contextual embeddings to improve relationship detection. Matentzoglu et al. [180] demonstrated how LLMs like MapperGPT enhance entity linking and mapping precision. Wu et al. [181] suggested an approach using online clustering with LLMs for ontology expansion.

Additionally, studies have explored applying LLMs for entity resolution, focusing on identifying, matching, and reconciling entities across different data sources. Hees [182] examined LLMs for aligning non-ontological data to ontologies, focusing on semantic interoperability in the energy sector. Using GPT-3.5 and GPT-4, the study aligned datasets from the ENERSHARE project with the Semantic TreeHouse ontology framework. The findings suggest that while LLMs support entity resolution, they function best as recommendation tools rather than deterministic alignment methods.

Zhang et al. [183] explored weakly-supervised taxonomy enrichment, demonstrating how LLMs can identify discriminative features to enhance hierarchical structures. Their study introduced self-supervised entity resolution techniques and seed-guided taxonomy construction, showing that minimal human supervision enables LLMs to build structured ontologies from text corpora.

Hofer et al. [184] analyzed LLM-driven entity reconciliation in KG construction, emphasizing the need for incremental ontology updates. Their findings indicate that while LLMs generate entity alignment candidates, a hybrid approach combining LLMs with rule-based validation is necessary for improving accuracy. The study highlights LLM applications in ontology alignment, particularly in biomedical KGs, where they refine taxonomic structures while maintaining semantic consistency.

Chen et al. [185] examined schema-based KGs for taxonomic reconciliation, leveraging OWL ontologies, RDFS, and SHACL. The study underscores the importance of ontology mapping repositories (e.g., UMLS Metathesaurus, BioPortal, MONDO) in resolving cross-domain taxonomic inconsistencies. Additionally, it explores LLM integration in entity resolution, particularly in instance matching and automated ontology extension.

These works emphasize the potential for LLMs to automate and improve ontology engineering, which is also investigated in this study. Table 7.1 summarizes how these works relate to our study.

Despite these developments, integrating LLMs into ontology engineering is challenging. Challenges related to model interpretability and reliability [174, 186, 187], domain specificity, and computational costs [188] are some of the key obstacles. Furthermore, the reliability of LLM-generated outputs in highly technical fields remains a concern, as hallucination and data bias continue to affect the consistency and accuracy of results [189]. Overcoming these challenges is crucial to guaranteeing the robustness and real-world applicability of LLM-supported ontology engineering.

7.3 Methods

Ontologies and KGs are closely related, with ontologies often serving as foundational frameworks for creating and interpreting KGs. Ontologies provide structured representations of domain-specific knowledge, including classes, relationships, and constraints. This logical foundation aligns with the principles of KGs, which are graph-based structures used to model interconnected data for reasoning and discovery [190, 191]. Ontologies not only aid in formalizing the semantics of KGs but also extend their capabilities by enabling advanced reasoning tasks [191]. Furthermore, they enhance KG construction by ensuring semantic interoperability and consistency, particularly when generating KGs from textual data [192]. The overlap between ontologies and KGs illustrates

Table 7.1 Relationship Between Related Studies and this Chapter

Study	Focus area	Relationship with this study
Aggarwal et al. [175]	LLMs for scholarly ontology generation	Automated concept and relationship extractions in ontology development
Zaitoun et al. [176]	LLMs for biomedical ontology augmentation	Demonstrates LLMs' ability to identify missing ontological relationships
Trajanoska et al. [177], Reis et al. [178]	LLMs for KG construction from unstructured text	LLM automation of ontology engineering tasks
Amini et al. [179]	LLMs for ontology alignment using contextual embeddings	Enhances taxonomic data alignment through contextual embeddings
Matentzoglu et al. [180]	LLMs for entity linking and mapping	Improves entity matching, crucial for taxonomic reconciliation
Wu et al. [181]	LLMs for ontology expansion using online clustering	Provides insights into LLM-based ontology expansion
Hees [182]	LLMs for semantic interoperability and ontology alignment	Applies LLMs to ontology alignment, relevant for aligning non-ontological data to an ontology
Zhang et al. [183]	LLMs for weakly-supervised taxonomy enrichment	Supports taxonomy enrichment with minimal human supervision
Hofer et al. [184]	LLMs for incremental ontology updates and entity reconciliation	Enhancing the scalability of ontology completion, updates, and creation using LLMs to extract and integrate new concepts from structured data, such as tabular datasets.
Chen et al. [185]	KGs for taxonomic reconciliation and LLM integration	Demonstrates how schema-based KGs aid taxonomic integration with LLMs

how these paradigms coalesce to support complex data representation and reasoning [193]. Given this close relationship, ontology development can benefit from recent techniques employed in KG construction, particularly applying LLMs.

The method employed in this case builds upon the framework outlined by [11] for LLM-augmented KG development to explore this synergy. The authors address different techniques using LLMs in KG development, including LLM-augmented KG Construction, and describe two approaches: end-to-end KG construction and distilling KGs from LLMs. According to [11], we can understand these approaches as follows:

- **End-to-end KG construction** is an innovative approach utilizing LLMs to streamline KG creation from raw text, bypassing the conventional multi-stage process [11]. Studies like [194] have developed methods integrating LLMs for named entity recognition and relation extraction, employing dual BERT-based classifiers to identify relationships and their orientations between entities [11]. This method allows for the direct assembly of KGs, showcasing models like Grapher and PiVE that facilitate KG generation from textual descriptions or correct errors in existing KGs using iterative verification with smaller LLMs [11].

- **Distilling KGs** from LLMs capitalizes on the inherent knowledge encoded within LLMs to extract and structure this information into KGs [11]. Techniques like COMET [195] leverage LLMs to generate commonsense KGs by training on seed sets of knowledge tuples, enabling the LLM to produce novel, high-quality knowledge tuples [11]. This approach underscores the potential of LLMs as tools for enhancing KG construction and as vast repositories of implicit knowledge ripe for extraction and organization into explicitly structured forms [11].

The methodological approach employed in this chapter is similar to the End-to-End KG Construction but less complex since we do not deal with extraction from raw text. The data input we used in the ontology construction is structured as JSON files with representative and precise names, which makes it easier to process (and more precise) than natural language input.

In this research, we used ChatGPT-4 to improve the quality of an ontology. Two alternative approaches were explored: (1) issuing a series of prompts for ChatGPT-4 to execute tasks via the BrowserOP plugin and (2) directing ChatGPT-4 to design a Python algorithm to perform analogous tasks. We followed the workflow activities shown in Fig. 7.1 in both approaches. The following sections discuss each activity in detail.

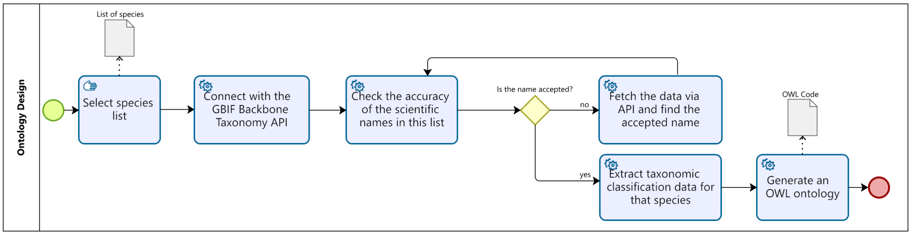

Ontology Design
List of species
Select species list
Connect with the GBIF Backbone Taxonomy API
Check the accuracy of the scientific names in this list
Is the name accepted?
no
yes
Fetch the data via API and find the accepted name
Extract taxonomic classification data for that species
OWL Code
Generate an OWL ontology

Figure 7.1 Workflow for the automated ontology design.

7.3.1 Selecting the Species List

We started by selecting a list of plant and animal species related to all agricultural product types in APTO. This step involved mapping product types from three key Brazilian datasets onto commodity price indexes, namely the agricultural prices index from CEPEA [122], the IpeaData Macroeconomic Indicators [124], and the CONAB agricultural prices database [126]. We based the association between product types and species on resources such as Agrotermos, Agrovoc, and the scientific literature, producing two lists: an Animal species list with 14 scientific names and a Plant species list with 74 names, both lists with names of species or genus level.

Another criterion for selecting these species was their agricultural relevance, particularly within Brazilian agronomic and livestock domains. The list includes economically significant species, widely cultivated, and well-documented in key Brazilian agricultural data repositories. Given Brazil's diverse agricultural landscapes, these species provide a representative yet incomplete view of the taxonomic groups involved in commercial agriculture.

While the selection is representative of Brazilian agrobiodiversity, it has limitations when generalizing to global taxonomies. Many agricultural taxonomies include species from other regions with different climatic, economic, and ecological contexts. As this study focuses on Brazilian datasets, certain species with regional importance elsewhere may be underrepresented. However, the approach developed in this study is not limited to Brazilian species, as it can be extended to include species from other regions, as long as they are represented in the GBIF Backbone Taxonomy.

7.3.2 Connecting with the GBIF Backbone Taxonomy API

GBIF "is an international network and data infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth" [196]. GBIF is recognised as the "[...] largest biodiversity data infrastructure in terms of the number of observations" [197], and supports a diverse range of stakeholders, from research scientists to governmental policy-makers, by aggregating biodiversity data from multiple sources, standardizing it, and making it freely available online [198].

For the taxonomy of organisms, we imported scientific names from the GBIF Backbone Taxonomy [199]. GBIF provides an API[1] for extracting taxonomic classifications. This API operates on data stored within the GBIF Checklist Bank, which indexes every registered checklist dataset across the GBIF network [200]. To consume these RESTful JSON web services, GBIF internally utilizes a Java web service client [200]. Using ChatGPT-4 and the BrowserOp plugin, we automated the retrieval of taxonomic data for each species via the GBIF API.

---

[1] `https://api.gbif.org/v1/`

The GBIF Backbone Taxonomy is a comprehensive synthetic classification system that underpins GBIF's integration of diverse name-based information from various sources, including occurrence datasets, species pages, and external databases like the Encyclopedia of Life (EOL), Genbank, and the International Union for Conservation of Nature and Natural Resources (IUCN) red list database, facilitating consistent taxonomic searches and cross-referencing across resources [199]. Regular updates expand its scope by merging additional scientific names from authoritative datasets into the Catalogue of Life-based structure[201], enhancing coverage beyond family-level classifications [199]. The taxonomy also incorporates identifiers for Operational Taxonomic Units (OTUs) from barcoding projects like iBOL and UNITE, employing consensus-based methods to connect Barcode Index Numbers (BINs) and Species Hypotheses (SHs) to taxon names and classifications [199]. Available for download in various formats, the GBIF Backbone Taxonomy data represents a critical resource for taxonomic research, drawing from over 105 sources to ensure broad and accurate name coverage [199].

7.3.3 Checking the Accuracy of Scientific Names

The assignment of scientific names to species is governed by international codes of nomenclature, which establish the principles and rules for naming organisms to ensure clarity and consistency across the scientific community. This process involves the formal description and publication of each species, adhering to specific criteria set by these codes. Over time, scientific names may become deprecated due to various factors, such as the discovery of prior valid names overlooked in earlier assessments or new phylogenetic insights that necessitate reclassification [202, 203]. As a result, the nomenclatural landscape is dynamic, with taxonomists revising and updating names to reflect the most accurate understanding of biological relationships. This leads to synonymizing many species names [202, 203]. This evolutionary nature of taxonomies underscores the importance of continual review and adherence to the international codes to maintain the integrity of scientific communication.

We reviewed every name on the species list to determine whether they were accepted or synonyms. For synonyms, the first approach involved using ChatGPT to retrieve the accepted name through the GBIF API. This process has also been implemented in the second approach using the Python Taxonomy Converter.

7.3.4 Extracting Multi-Level Taxonomic Classification Data

Species classification follows various conventions that lead to multi-level taxonomies that organize life forms into hierarchical categories

[204]. In this study, we adopted a seven-tiered framework for the taxonomic classification of species, which includes kingdom, phylum, class, order, family, genus, and species, aligning with established biological taxonomy principles [204]. This hierarchical structure is crucial for understanding the relationships and distinctions among different life forms [204]. During the classification phase, we tasked both applications (ChatGPT integrated with the BrowserOp plugin and the Python Taxonomy Converter) with extracting these seven levels of classification for every species listed or six levels for names at genus level.

7.3.5 Generating an OWL Ontology

Finally, we used the extracted taxonomic information to generate an OWL ontology, modeling the information as classes and subclasses in OWL/XML, complete with labels and URIs sourced from GBIF.

7.4 Approach 1: ChatGPT with BrowserOp Plugin

7.4.1 Prompt Specification and Results

The process consisted of incorporating a series of instructions into the ChatGPT 4 model, as shown in Table 7.2. The prompt engineering methodology used in this study builds on Giray's work [205], which categorizes the components of an effective prompt into four main elements:

- **Instruction**: A directive that specifies the task, guiding the model's behavior and ensuring it aligns with the desired outcome.
- **Context**: Relevant background information or external details provide additional knowledge, enabling the model to generate accurate and context-aware responses.
- **Input data**: The core query or data to be processed serves as the primary driver of the model's understanding and response generation.
- **Output indicator**: An indicator of the expected response format, such as a brief answer, detailed explanation, or structured output, shaping the model's solution delivery.

We built the prompt based on these four elements, as shown in Table 7.2.

The species names used as **Input data** are listed in Table 7.3. The first species in this list, *Apis mellifera*, served as a test case where the ChatGPT model utilized the BrowserOp plugin to send a request to the GBIF API. This plugin fetched real-time data from the web [206], and facilitated the interaction as detailed in Listing 7.1. As specified in **Context**, the call to the GBIF API performed with the BrowserOP plugin

Table 7.2 Prompt Specification

Agent	Text
👤	Task: Build an ontology for species names. Step 0: Access the GBIF API: `https://api.gbif.org/v1/` Step 1: Extract the full taxonomic classification for each name in this list, including kingdom, phylum, class, order, family, genus, and species. In some cases, the name on the list is at the species level, and in other cases, at the genus level. Step 2: This taxonomy should be organized as OWL classes and subclasses (do not create individuals or instances). Each class should have a URI from GBIF and an RDF label with the name for that class, with the language tag (considering that scientific names are written in Latin, include the language tag ”lat”). Step 3: if a name has the status ”deprecated”, look for the accepted name for that species. Step 4: output: generate an OWL code, including all the necessary namespaces.

successfully retrieved the desired data. In our directive to ChatGPT, we specified a connection with the GBIF API’s base URL (`https://api.gbif.org/v1/`), which hosts various endpoints catering to different data dimensions, including but not limited to the Registry API, Species API, Occurrence API, and Literature API, each serving distinct purposes such as registry lookup, species matching, occurrence data retrieval, and literature references, respectively. Despite the lack of explicit instruction on which endpoint to use, ChatGPT identified and targeted the appropriate API endpoint for species matching, which is `https://api.gbif.org/v1/species/`.

Table 7.3 List of Animal Scientific Names

Taxon rank	Taxon name
Species	*Apis mellifera*
Species	*Bos taurus*
Subspecies	*Capra aegagrus hircus*
Species	*Ovis aries*
Genus	*Sus*
Species	*Bombyx mori*
Species	*Prochilodus cearensis*
Species	*Prochilodus scrofa*
Species	*Prochilodus margravii*
Species	*Semaprochilodus insignis*
Species	*Semaprochilodus taeniunes*
Species	*Colossoma mitrei*
Species	*Arapauma gigas*
Genus	*Tilapia*

Listing 7.1 Taxonomy Extraction

```
// Request to BrowserOP
{
  "link": "https://api.gbif.org/v1/species/match?name=Apis%20mellifera"
}

// Response from BrowserOp
{
  "result": {
    "content": "{\"usageKey\":1341976,\"scientificName\":\"Apis
          ↪ mellifera Linnaeus, 1758\",\"canonicalName\":\"Apis mellifera
          ↪ \",\"rank\":\"SPECIES\",\"status\":\"ACCEPTED\",\"confidence
          ↪ \":99,\"matchType\":\"EXACT\",\"kingdom\":\"Animalia\",\"
          ↪ phylum\":\"Arthropoda\",\"order\":\"Hymenoptera\",\"family\":
          ↪ \"Apidae\",\"genus\":\"Apis\",\"species\":\"Apis mellifera\",
          ↪ \"kingdomKey\":1,\"phylumKey\":54,\"classKey\":216,\"orderKey
          ↪ \":1457,\"familyKey\":4334,\"genusKey\":1334757,\"speciesKey
          ↪ \":1341976,\"synonym\":false,\"class\":\"Insecta\"}\n",
    "rules": [
      "Use the language that the user previously used or the language
          ↪ requested by the user.",
      "Based on the provided data (and all historical retrieved data if
          ↪ the results need more than 1requests to be completely
          ↪ provided), respond to the user's request. No need to
          ↪ include other irrelevant words in your answer.",
      "Always extract url or source from provided content as a support
          ↪ for your answer.",
      "Ensure the information provided is accurate, clear, and concise.
          ↪ Avoid overly complex or lengthy explanations when a simple,
          ↪  direct response will suffice, while in the meantime, make
          ↪ sure to provide all the information the user will find
          ↪ valuable based on your correct understanding of user's
          ↪ intention.",
      "If a conclusive answer cannot be generated based on the provided
          ↪ content, respond to user's query first and then attempt to
          ↪ interpret the user's intent and suggest ways for them to
          ↪ refine their query for a more successful result.",
      "If the provided data includes authoritative source such as [wiki]
          ↪  or [official website], organize the response based
          ↪ primarily on data from these sources.",
      "If the user's query contains a URL, give a brief description of
          ↪ the site and summarize it based on the content of the page
          ↪ ."
    ]
  },
  "code": 0
}
```

The API response shown in Listing 7.1 adheres to the prompt specification (refer to Table 7.2), demonstrating the successful extraction of detailed taxonomic classification data for the species *Apis mellifera*. As specified in the prompt's **Input data**, ChatGPT retrieved the complete taxonomic hierarchy, including the ranks kingdom, phylum, class, order, family, genus, and species. Listing 7.1 confirms the accurate extraction of all ranks, formatted as key-value pairs (e.g., kingdom:Animalia).

The prompt further specified that the ontology generated from the retrieved (**Output indicator** in Table 7.2) data should incorporate GBIF

URIs for each class representing a taxon. Listing 7.1 shows that each taxon is assigned a unique `key` corresponding to its taxonomic rank (e.g., kingdomKey:1, genusKey:133475). GBIF assigns a unique numeric ID to each entity in its Backbone Taxonomy, forming the URI that uniquely identifies every scientific name in its catalog. For example, the response indicates Animalia is linked to the `kingdomKey` value 1, which is represented by the GBIF URI `https://www.gbif.org/species/1`. These URIs ensure that each taxon is unambiguously identified within the ontology.

The prompt also required ChatGPT to verify the status of the species' name. The API response in Listing 7.1 shows the name status as `ACCEPTED`, indicating that the species name is valid and recognized within the taxonomy catalog. This outcome confirms ChatGPT's successful execution of this verification step.

Moreover, Listing 7.1 shows the `rules` block, which outlines a set of operational guidelines not explicitly defined in the prompt, suggesting they originate from ChatGPT's model internal rules to enhance the response's comprehensiveness and relevance.

Finally, the prompt in Table 7.2 defined that the output should be an OWL file generated from the data extracted via the API. Initially, GPT-4 created OWL code using `https://api.gbif.org/v1/species/` as the namespace for the ontology classes. This URI is a reference point within the GBIF API, linking to JSON data files describing scientific names and their relationships, as illustrated in Listing 7.2.

Listing 7.2 JSON Data for the Animalia Class (https://api.gbif.org/v1/species/1)

```
{
  "key": 1,
  "nubKey": 1,
  "nameKey": 130188353,
  "taxonID": "gbif:1",
  "sourceTaxonKey": 1,
  "kingdom": "Animalia",
  "kingdomKey": 1,
  "datasetKey": "d7dddbf4-2cf0-4f39-9b2a-bb099caae36c",
  "constituentKey": "d7dddbf4-2cf0-4f39-9b2a-bb099caae36c",
  "scientificName": "Animalia",
  "canonicalName": "Animalia",
  "vernacularName": "Animals",
  "authorship": "",
  "nameType": "SCIENTIFIC",
  "rank": "KINGDOM",
  "origin": "SOURCE",
  "taxonomicStatus": "ACCEPTED",
  "nomenclaturalStatus": [],
  "remarks": "",
  "numDescendants": 2981931,
  "lastCrawled": "2023-08-22T23:20:59.545+00:00",
  "lastInterpreted": "2023-08-22T22:11:51.237+00:00",
 "issues": []
}
```

However, after careful consideration, we decided to use URIs from the `https://www.gbif.org/species/` endpoint instead. We chose this endpoint because it provides access to valuable enriched resources for end users, such as taxon photos, occurrence statistics, and related information. Although we did not import these supplementary resources into the ontology, users can access them by clicking on the URL, aligning with linked data principles.

To apply this change, we directed ChatGPT to regenerate the OWL code using `https://www.gbif.org/species/` as the namespace. Listing 7.3 shows the updated output.

Listing 7.3 GPT-Generated OWL Representation for Taxonomic Hierarchy

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
       xmlns:owl="http://www.w3.org/2002/07/owl#">

   <!-- Kingdom Animalia -->
   <owl:Class rdf:about="https://www.gbif.org/species/1">
      <rdfs:label xml:lang="lat">Animalia</rdfs:label>
   </owl:Class>

   <!-- Phylum Arthropoda -->
   <owl:Class rdf:about="https://www.gbif.org/species/54">
      <rdfs:label xml:lang="lat">Arthropoda</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/1"/>
   </owl:Class>

   <!-- Class Insecta -->
   <owl:Class rdf:about="https://www.gbif.org/species/216">
      <rdfs:label xml:lang="lat">Insecta</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/54"/>
          ↪
   </owl:Class>

   <!-- Order Hymenoptera -->
   <owl:Class rdf:about="https://www.gbif.org/species/1457">
      <rdfs:label xml:lang="lat">Hymenoptera</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/216"/
          ↪ >
   </owl:Class>

   <!-- Family Apidae -->
   <owl:Class rdf:about="https://www.gbif.org/species/4334">
      <rdfs:label xml:lang="lat">Apidae</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
          ↪ 1457"/>
   </owl:Class>

   <!-- Genus Apis -->
   <owl:Class rdf:about="https://www.gbif.org/species/1334757">
      <rdfs:label xml:lang="lat">Apis</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
          ↪ 4334"/>
   </owl:Class>

   <!-- Species Apis mellifera -->
   <owl:Class rdf:about="https://www.gbif.org/species/1341976">
      <rdfs:label xml:lang="lat">Apis mellifera</rdfs:label>
      <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
          ↪ 1334757"/>
   </owl:Class>

</rdf:RDF>
```

We repeated this process for all taxa listed in Table 7.3. ChatGPT confirmed the taxon names *Bos taurus*, *Ovis aries*, Genus *Sus*, *Semaprochilodus insignis*, *Semaprochilodus taeniurus*, *Arapaima gigas*, and *Tilapia* as accepted names within the GBIF Backbone Taxonomy. Then, it successfully generated the corresponding OWL code for these taxa. The OWL files are available on Zenodo [10].

ChatGPT identified the remaining taxa as having outdated or deprecated names, classifying them as synonyms, as shown in Table 7.4. These taxon names were listed as preferred terms in the Agrotermos thesaurus but were absent from AGROVOC, the two vocabularies used as the foundation for building this ontology. An exception is *Capra hircus*, which we did not find in either of these vocabularies.

Table 7.4 Deprecated Scientific Names (Synonyms) and Corresponding Accepted Names Detected by ChatGPT.

Target name	Synonym of
Capra aegagrus hircus	*Capra aegagrus*
Prochilodus cearensis	*Prochilodus brevis*
Prochilodus scrofa	*Prochilodus lineatus*
Prochilodus margravii	*Prochilodus argenteus*
Colossoma mitrei	*Piaractus mesopotamicus*

Agrotermos and AGROVOC include concepts for representing ‘goat’, but with limitations. In Agrotermos, the term ‘Cabra’ (Portuguese for goat) is not explicitly associated with its scientific species. In AGROVOC, *Capra hircus* is modeled as an entry term for this concept. To address this gap, we referred to the classification provided on Wikipedia, which defines ‘goat’ species as *Capra hircus*. We added this name to the list, which turned out to be outdated.

ChatGPT retrieved the corresponding accepted name from the GBIF Backbone Taxonomy for each synonym identified and generated OWL code incorporating these validated names, as shown in Listing 7.4.

Listing 7.4 Iterations to Fix Incorrect and Outdated Information (Synonyms)

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
```

```
    xmlns:owl="http://www.w3.org/2002/07/owl#">

<!-- Kingdom Animalia -->
<owl:Class rdf:about="https://www.gbif.org/species/1">
   <rdfs:label xml:lang="lat">Animalia</rdfs:label>
</owl:Class>

<!-- Phylum Chordata -->
<owl:Class rdf:about="https://www.gbif.org/species/44">
   <rdfs:label xml:lang="lat">Chordata</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/1"/>
</owl:Class>

<!-- Class Actinopterygii -->
<owl:Class rdf:about="https://www.gbif.org/species/204">
   <rdfs:label xml:lang="lat">Actinopterygii</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/44"/>
       ↪
</owl:Class>

<!-- Order Characiformes -->
<owl:Class rdf:about="https://www.gbif.org/species/537">
   <rdfs:label xml:lang="lat">Characiformes</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/204"/
       ↪ >
</owl:Class>

<!-- Family Prochilodontidae -->
<owl:Class rdf:about="https://www.gbif.org/species/7331">
   <rdfs:label xml:lang="lat">Prochilodontidae</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/537"/
       ↪ >
</owl:Class>

<!-- Genus Prochilodus -->
<owl:Class rdf:about="https://www.gbif.org/species/2352148">
   <rdfs:label xml:lang="lat">Prochilodus</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
       ↪ 7331"/>
</owl:Class>

<!-- Species Prochilodus brevis (for Prochilodus cearensis) -->
<owl:Class rdf:about="https://www.gbif.org/species/2352151">
   <rdfs:label xml:lang="lat">Prochilodus brevis</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
       ↪ 2352148"/>
</owl:Class>

<!-- Species Prochilodus lineatus (for Prochilodus scrofa) -->
<owl:Class rdf:about="https://www.gbif.org/species/2352154">
   <rdfs:label xml:lang="lat">Prochilodus lineatus</rdfs:label>
   <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/
       ↪ 2352148"/>
</owl:Class>

<!-- Species Prochilodus argenteus (for Prochilodus margravii) -->
```

```
  <owl:Class rdf:about="https://www.gbif.org/species/2352177">
     <rdfs:label xml:lang="lat">Prochilodus argenteus</rdfs:label>
     <rdfs:subClassOf rdf:resource="https://www.gbif.org/species/2352148"/>
  </owl:Class>

</rdf:RDF>
```

In Listing 7.4, we tasked ChatGPT with generating OWL code for three species within the same genus: *Prochilodus Cearensis*, *Prochilodus Scrofa*, and *Prochilodus Margravii*. However, during our first attempt to extract data from the GBIF API using these names, ChatGPT reported that no data was available for these taxa. We discovered that this issue stemmed from formatting errors in the species names. According to international standards for species nomenclature [157, 158], species names should consist of a capitalized genus name followed by a lowercase specific epithet, often including the naming authority and year of identification. The names we used in the initial query, copied directly from the Agrotermos vocabulary, had capitalization errors in the specific epithets (e.g., *Cearensis* instead of *cearensis*), which rendered them unrecognizable to the GBIF API.

After correcting the formatting to align with these standards, we reissued the query, and ChatGPT successfully retrieved the taxonomic data. Table 7.5 provides a detailed log of this interaction, highlighting the issue and the resolution process.

Table 7.5 Iteration with ChatGPT

Agent	Text
👤	Do it [repeat the steps described in Table 7.2] now for these three related species: Prochilodus Cearensis, Prochilodus Scrofa, and Prochilodus Margravii.
֍	The attempts to match the species names “Prochilodus Cearensis,” “Prochilodus Scrofa,” and “Prochilodus Margravii” with the GBIF database did not yield any results due to too little confidence in the matches. This means that the specific names provided might not be directly available in the GBIF database under those exact spellings, or are not recognized at the species level in the database. For taxonomic research and ontology building, accurate species names are crucial. It is possible that these names might have synonyms under which they are listed in databases, or there might be slight variations in spelling or taxonomy that affect the search results. In such cases, further steps could include: Verifying the scientific names for spelling or taxonomic updates. Searching for broader taxonomic groups (e.g., the genus “Prochilodus”) and then narrowing down to the species level based on additional sources or databases. Consulting specialized taxonomic databases or literature for the most current classification and synonyms. If you have alternative or broader taxonomic categories in mind or need assistance with another query, please let me know how I can assist further.

Agent	Text
👤	Try again with: Prochilodus cearensis, Prochilodus scrofa, and Prochilodus margravii.
֍	Based on the data retrieved from the GBIF API for the species Prochilodus cearensis, Prochilodus scrofa, and Prochilodus margravii, here's the ontology in OWL format. It is important to note that all three names are recognized as synonyms for other species within the GBIF database: [provided the OWL code shown in Listing 7.4. After the code block, it commented:] This OWL code represents the full taxonomic classification for the species Prochilodus cearensis (synonym for Prochilodus brevis), Prochilodus scrofa (synonym for Prochilodus lineatus), and Prochilodus margravii (synonym for Prochilodus argenteus), including the higher taxon levels (kingdom, phylum, class, order, family, genus) with each class having a valid URI from GBIF and a rdf label of the name for that class in Latin.

Following the initial incident, we intentionally repeated the same mistake by submitting the species name *Colossoma Mitrei* with the same capitalization error in the specific epithet. ChatGPT corrected the error autonomously before sending the request to the GBIF API, demonstrating its ability to learn iteratively from prior interactions.

This adaptability reflects ChatGPT's advanced learning capabilities, allowing it to refine its comprehension and response strategies over time. Such functionality exemplifies a core characteristic of AI: the ability to internalize lessons from previous interactions and apply this knowledge to similar situations in the future, thereby reducing errors and enhancing operational efficiency [207, 208].

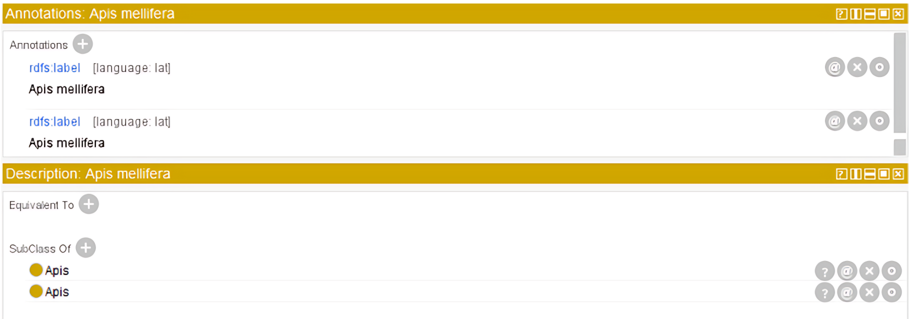
In ontology development, this AI-driven error correction potentially reduces manual effort and enhances data quality. Studies [207, 208, 209, 210] highlight the growing role of AI in fostering intelligent and efficient system interactions, making it a transformative tool in fields that demand adaptability and precision.

7.4.2 Limitations

This approach presented limitations when handling extensive species lists. For this reason, we did not move forward with generating the OWL code for the plant species listed in the ontology since the list of plant species is larger than the animal species list.

Moreover, the processing time increases with the list's size, impacting this method's scalability. For example, the processing time to call the GBIF API and generate the ontology for the species *Apis mellifera* was 37 seconds. For the three species *Prochilodus cearensis*, *Prochilodus scrofa*, and *Prochilodus margravi*, which were synonyms, the processing time was 2 minutes and 5 seconds. This approach proved efficient for smaller clusters of species names, particularly when the names are taxonomically related (within the same genus or family). However, we observed that reliability decreases for lists with more than five names, where the likelihood of hallucinations increases. Initially, we attempted to process the entire list of species names from Table 7.3 through ChatGPT in a single batch. However, this approach resulted in multiple inaccuracies, including the generation of URIs non-existent in the GBIF database, incorrect association between valid URIs and labels, and occasional omissions of tasks, such as adding RDF labels. We opted for a more controlled approach to address these issues by submitting names to ChatGPT individually or in smaller groups (at most three species each time). This strategy significantly improved the accuracy of the results.

Another challenge was integrating multiple OWL files using Protégé, which resulted in duplicated labels and parent-child relationships as illustrated in Fig. 7.2. Despite attempts to consolidate these files using Protégé's merge function, these redundancies persisted, but this can be also considered as a limitation of Protégé for handling these multiple files rather than a problem with ChatGPT-generated code.



Annotations: Apis mellifera

Annotations

rdfs:label [language: lat]
Apis mellifera

rdfs:label [language: lat]
Apis mellifera

Description: Apis mellifera

Equivalent To

SubClass Of
Apis
Apis

Figure 7.2 Duplicated Labels and Relationships in the Ontology Visualization in Protégé.

Another limitation of this approach is its poor reproducibility due to ongoing updates in ChatGPT's technology. During the initial experiments, ChatGPT plugins such as BrowserOP were still operational. However, as of March 2024, OpenAI replaced these plugins with GPTs, and ChatGPT no longer has direct access to the Internet.

To assess the feasibility of reproducing the results, we attempted to replicate the experiments on March 6, 2025, using three different GPT models: BrowserOP (the former plugin), ChatGPT-4o, and Data Analyst. In all three cases, the models returned the same message indicating their inability to make direct API calls, confirming that ChatGPT currently lacks Internet access. Therefore, this approach is no longer viable unless ChatGPT restores internet access in the future.

To evaluate whether a similar approach could work with alternative AI models, we tested the same prompt using DeepSeek V3 and Google Gemini 2.0. However, the results were highly inaccurate and unreliable. Although both models generated API responses that structurally resembled the expected format (similar to Listing 7.1), the responses contained incorrect taxon keys that did not match the expected labels. Additionally, neither model correctly identified that the species in this test (*Prochilodus scrofa*) was a synonym.

Even after multiple refinements to the prompt, including providing more details about API syntax and sample responses, both models continued to generate hallucinated results that did not accurately reflect the expected outputs. These issues indicate that using this approach with DeepSeek V3 and Gemini 2.0 would require further fine-tuning or prompt engineering to mitigate these errors. The full prompts and interaction transcripts with DeepSeek and Google Gemini are available on Zenodo [10].

In conclusion, this approach is currently not reproducible with ChatGPT due to the lack of internet access. Additionally, attempts to apply this methodology to other AI models, such as DeepSeek V3 and Gemini 2.0, require further adjustments to reduce hallucinations and improve reliability. Therefore, we recommend using Approach 2 as a more viable solution for generating an OWL ontology for species taxonomy.

7.5 Approach 2: Python Algorithm for Converting GBIF Taxonomic Data to OWL

Building on the steps outlined in Table 7.2 and the prompt interactions described in Section 7.4, we tasked ChatGPT with developing a Python algorithm to generate OWL files from data retrieved via the GBIF Backbone Taxonomy API. However, the initial version of the algorithm introduced redundancies in the generated OWL code by repeating higher taxonomic levels for each taxon. For example, the class *Animalia* was redefined for every taxon describing an animal species rather than being defined once and used as a superclass for the relevant species. To address this issue, we provided ChatGPT with the revised instruction outlined in Table 7.6.

Table 7.6 Iteration with ChatGPT: Fixing Repetitions

Agent	Text
👤	It should present the superior taxon levels just once in the code when they are the same for multiple species. Animalia, for instance, is a taxon common to all species in this list.
❀	To address the issue of repeating higher taxon levels for multiple species, you can modify the script to accumulate unique taxon levels across all species before generating the OWL representation. This approach ensures that each taxon level (e.g., kingdom, phylum) is represented only once, regardless of how many species share it.

The algorithm second version fixed redundant taxonomic level definitions but introduced a new issue: improper class-subclass relationships in the OWL code. When reported to ChatGPT, it responded as shown in Table 7.7.

Table 7.7 Iteration with ChatGPT: Fixing Relationships

Agent	Text
	To ensure the OWL output correctly represents the `subClassOf` relationships and avoids duplicating common higher taxon levels across species, let’s refine the approach. We’ll adjust the script to build a hierarchy that maintains these relationships and ensures each taxon level is included only once.

In this third trial, we got the error: “AttributeError: ‘list’ object has no attribute ‘replace,’ which was caused by a typo. We tasked ChatGPT to review the species list, and it detected *Semaprochilodus taeniunes* as a typo, suggesting the correct name as *Semaprochilodus taeniurus*. ChatGPT inference was correct since this is also the accepted name in the GBIF Backbone Taxonomy. After reviewing this name and reentering the list, the algorithm worked as expected.

Finally, we tasked ChatGPT with generating the documentation for the algorithm, which we named Taxonomy OWLizer. The complete algorithm and its documentation are available on Zenodo [11].

7.5.1 Script Overview

This Python script follows a pipeline to fetch taxonomic classification data for a species list from the GBIF API and generates an OWL representation. It can handle multiple species names, representing higher taxonomic levels only once, even if they are shared across species, avoiding duplication of classes as reported in the limitations of Approach 1. It also checks synonyms and looks for the accepted names. The script is written in Python 3.x and requires the `requests` library.

Functions

The code generated by ChatGPT attended the requirements specified in Table 7.2. The following functions were included:

- `fetch_gbif_data(scientific_name)`: this function fetches taxonomic classification from the GBIF API for a given scientific name.
- `fetch_synonyms(species_key)`: Fetches synonyms for a given species key from the GBIF API. This is useful for verifying the accepted scientific name of a species and ensuring that the OWL ontology only includes the correct names.
- `accumulate_taxa(data, taxa)`: accumulates unique taxa and their relationships from the GBIF data.

- `validate_accepted_name(data, taxa)`: Ensures that only the accepted scientific name is included in the taxa dictionary by validating against synonyms.
- `generate_owl(taxa)`: generates an OWL/XML representation from the accumulated taxa.

Listing 7.5 shows an example of usage for the list of names from Table 7.3. The output was an OWL/XML file (`Animalia.xml` on Zenodo [10]). Subsequent inspection of this file in Protégé revealed a well-structured class hierarchy, depicted in Fig. 7.3. This visualization confirmed the accuracy of the hierarchical relationships and the appropriate formatting of the generated OWL code.

Listing 7.5 Example of Usage

```
species_names =[
   "Apis mellifera", "Bos taurus", "Capra aegagrus hircus", "Ovis
       ↪ aries", "Sus", "Bombyx mori", "Prochilodus cearensis", "
       ↪ Prochilodus scrofa", "Prochilodus margravii", "
       ↪ Semaprochilodus insignis", "Semaprochilodus taeniurus", "
       ↪ Colossoma mitrei", "Arapaima gigas", "Tilapia"
]

taxa ={}
for name in species_names:
   data =fetch_gbif_data(name)
   if data:
      accumulate_taxa(data, taxa)
      validate_accepted_name(data, taxa)

owl_data =generate_owl(taxa)
print(owl_data)
```

Testing the Algorithm with a Larger List of Species Names

The scalability limitation of Approach 1 prevented the generation of OWL codes for long species lists. In contrast, the second approach works significantly faster. To evaluate its efficiency, we conducted a test with 74 plant species listed in the file `PlantSpeciesList.py` available on Zenodo [10]. As an initial step, we asked ChatGPT to verify the spelling of the species names in this list using only its internal knowledge, without connecting to the GBIF API.

Although we did not explicitly request ChatGPT to identify synonyms since our algorithm handles this functionality, it flagged *Euterpe edulis* as a synonym of *Euterpe oleracea*. However, subsequent verification using authoritative sources, including the GBIF Backbone Taxonomy [199] and the Plants of the World Online catalog [211], confirmed that both *Euterpe edulis* and *Euterpe oleracea* are valid and accepted species names. This misidentification highlights the potential for inaccuracies

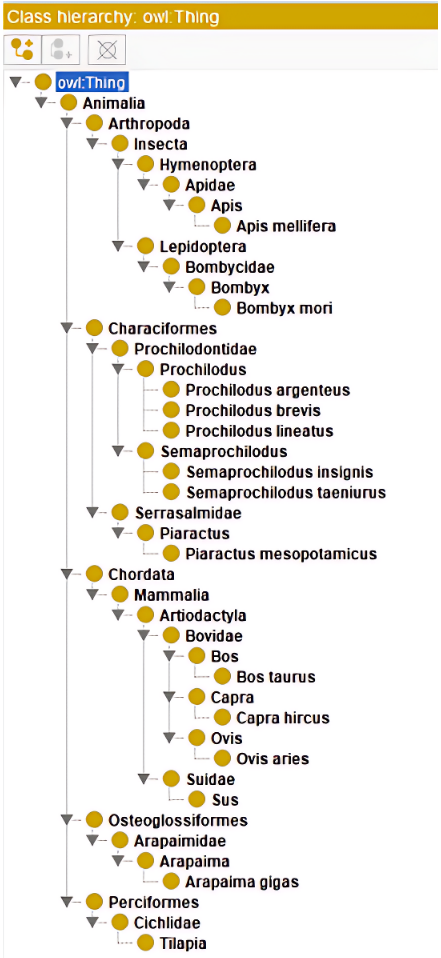


Figure 7.3 Classes Hierarchy on Protégé

when ChatGPT's analysis is based solely on its internal knowledge rather than being anchored to reliable external sources for validating taxonomic names.

Despite the incorrect synonym suggestion, we continued with the task. The algorithm performed efficiently, generating OWL code for nearly all the listed species (except for *Triticum x Secale*) in just 2 minutes and 31 seconds. By comparison, the first approach required 2 minutes and 5 seconds to generate OWL code for only three species. This stark contrast highlights the superior scalability of the second approach, which successfully processed 74 species at a similar time. The complete OWL code can be found as `plantae.xml` on Zenodo [10].

Hybrids in Ontological Modeling

The exclusion of *Triticum* × *Secale* from the output was due to inconsistencies in hybrid name formatting, which can vary across taxonomic catalogs. Hybrid species names are structured to indicate their origin from two different species or genera. While many taxonomic authorities use the symbol × to denote hybrid status – following conventions in botanical and zoological nomenclature – others may omit it [157, 212]. For interspecific hybrids, the × is placed between the genus name and the specific epithet; for intergeneric hybrids, it appears between the genus names [212]. For example, the intergeneric hybrid between Triticum (wheat) and Secale (rye) is typically written as *Triticum* × *Secale*.

In the GBIF Backbone Taxonomy – which aggregates names from various taxonomic sources – the use of the hybrid symbol is inconsistent. For instance, *Triticum secale*[2] appears without the ×, while *Citrus ×aurantium*[3] includes the symbol but omits the space. Notably, the GBIF API supports both formats when queried. We tested the algorithm using both *Triticum ×secale* (with the symbol and lowercase epithet) and *Triticum secale* (without the symbol), and both returned results, indicating that the name is recognized as a synonym of *Secale cereale*. To avoid issues during processing, hybrid names should be preprocessed into one of these formats before being submitted to the algorithm.

To enhance the expressiveness of hybrid modeling in our ontology, we manualy modeled hybrid relationships explicitly using an object property is_a_hybrid_of. For instance, the triple:

```
Citrus aurantium is_a_hybrid_of some (Citrus maxima and Citrus
    ↪ reticulata)
```

captures the hybrid lineage of *Citrus aurantium*. A corresponding OWL representation is shown in Listing 7.6, where the IRI https://www.gbif.org/species/8077391 identifies the class *Citrus aurantium*.

Listing 7.6 Modeling of hybrids in APTO

```
      <Class IRI="https://www.gbif.org/species/8077391"/>
      <ObjectSomeValuesFrom>
         <ObjectProperty IRI="is_a_hybrid_of"/>
         <ObjectIntersectionOf>
            <Class IRI="Citrus_maxima"/>
            <Class IRI="Citrus_reticulata"/>
         </ObjectIntersectionOf>
      </ObjectSomeValuesFrom>
```

The is_a_hybrid_of object property formally captures the hybridization relationship, linking the hybrid species class to the classes

---

[2] https://www.gbif.org/species/4154790

[3] https://www.gbif.org/species/8077391

representing its parent species. This approach provides a precise and machine-interpretable way to represent genetic lineage within the ontology.

7.5.2 Web Application with User Interface

Building on the Python script of Approach 2, we developed the Taxonomy OWLizer app[4], a web-based application that allows users to convert species names into an OWL taxonomy using the GBIF API. It provides a simple interface for users to enter species names, fetch their taxonomic classification, and download the generated OWL file. This tool is designed for biologists, taxonomists, and ontology researchers needing structured taxonomy data in OWL format.

The application supports the following key features:

1. Users can input the names of one or multiple species.
2. The app queries the GBIF API to fetch taxonomic classifications.
3. The app converts the taxonomy data into an OWL ontology format.
4. Users can copy and paste or download the generated OWL taxonomy file.
5. Previously searched species are cached in local storage, allowing instant retrieval without re-fetching data.

To develop the OWLizer, we utilized the following technologies:

- HTML, CSS (Bootstrap): Provides a simple and responsive user interface.
- JavaScript: Handles API requests, caching, and file downloads.
- jQuery: Simplifies DOM manipulation and AJAX requests.
- Flask (Python Back-end on Render): Processes API requests and fetches taxonomic data from GBIF.
- GBIF API: Provides taxonomic classification data.
- GitHub Pages: Hosts the front-end (static website) for free.
- Render (Free Tier): Hosts the back-end Flask API that fetches data.
- UptimeRobot: Sends periodic requests to prevent the back-end from going to sleep.

7.5.3 Limitations

Approach 2 successfully addressed the scalability limitations of Approach 1. However, this algorithm remains sensitive to typographical errors in the input data. To mitigate this, we recommend a preliminary

---

[4] `https://filipi-soares.github.io/TaxonOWLConverter/`

step in which users utilize ChatGPT to review the list of species names and correct typos. This workflow combines the strengths of both tools: ChatGPT's ability to verify and correct spellings, and the algorithm's efficiency in processing species names and performing synonym checks.

While the Taxonomy OWLizer web-based app provides a functional and accessible way to generate OWL representations of taxonomic data, it has certain limitations due to free-tier services. The back-end, which is hosted on Render's free tier, enters a sleep state after 15 minutes of inactivity, causing delays in API responses when it needs to restart. To mitigate this, UptimeRobot periodically pings the back-end, keeping it active most of the time. However, this workaround is not entirely reliable and may still result in occasional downtime. Since GitHub Pages is used to host the front-end, the application can only serve static content, requiring an external back-end to handle API requests.

Furthermore, the app caches responses using local storage, meaning that previously fetched data is only available on the same device and browser. This limits persistent caching across multiple sessions or devices, which could be improved by integrating a more robust storage solution such as IndexedDB or a cloud-based database. Despite these limitations, the current implementation remains an effective and free-to-use tool for taxonomists and ontology researchers.

7.6 Integrating the `:Organism` and `:Product_type` Modules

After generating and integrating the OWL file with species classifications into APTO Organism module, each product type was linked to its corresponding species. This association is essential for accurately representing the meaning of commercial product names, particularly within the regional context where they are traded. Establishing these precise connections ensures semantic clarity and helps prevent errors arising from ambiguous or incorrect associations.

APTO offers several examples that underscore the importance of this type of modeling. For example, a frequent ontological modeling error appears in sources such as Agrotermos, Agrovoc, and even in Wikipedia, concerning the term 'Pimenta.' In Brazilian Portuguese, 'Pimenta' is often translated into English as 'pepper.' While this translation might seem correct, it oversimplifies the term and creates semantic confusion. In Brazilian Portuguese, 'Pimenta' refers to the reddish or yellowish (sometimes green when collected unripe) berries from plants of the genus *Capsicum*. Examples include 'Pimenta malagueta' (*Capsicum frutescens*), 'Pimenta cumari' (*Capsicum baccatum var. praetermissum*), and 'Pimenta dedo-de-moça' (*Capsicum baccatum*).

In resources such as Agrovoc [213] and Agrotermos [214], the concept of 'Pimenta' is translated into English as 'pepper' and associated with the species *Piper nigrum*. In Brazil, however, *Piper nigrum* refers explicitly to the plant that produces black pepper, known locally as 'Pimenta-do-reino.'

In Agrotermos, *Capsicum* sp. is listed as a related concept to 'Pimenta,' but there is also a relationship indicating that 'Pimenta' is produced by *Piper nigrum*. This misrepresentation creates confusion, as the term 'Pimenta' in Brazil is predominantly associated with *Capsicum* sp., while 'Pimenta-do-reino' refers to *Piper nigrum* (black pepper), as previously mentioned.

Despite being a common mistake, the organisms represented by these concepts are biologically distinct. Table 7.8 shows the biological classification of both species, highlighting their differences. *Capsicum* sp. and *Piper nigrum* are only related at the class level (*Magnoliopsida*), underscoring their distinct evolutionary and biological characteristics.

Table 7.8 Classification of *Piper nigrum* and *Capsicum* sp. (Based on the GBIF Backbone Taxonomy [199]).

Taxon rank	**Piper nigrum**	**Capsicum sp.**
Kingdom	Plantae	Plantae
Phylum	Tracheophyta	Tracheophyta
Class	Magnoliopsida	Magnoliopsida
Order	Piperales	Solanales
Family	Piperaceae	Solanaceae
Genus	*Piper* L.	*Capsicum* L.
Species	*Piper nigrum* L.	—

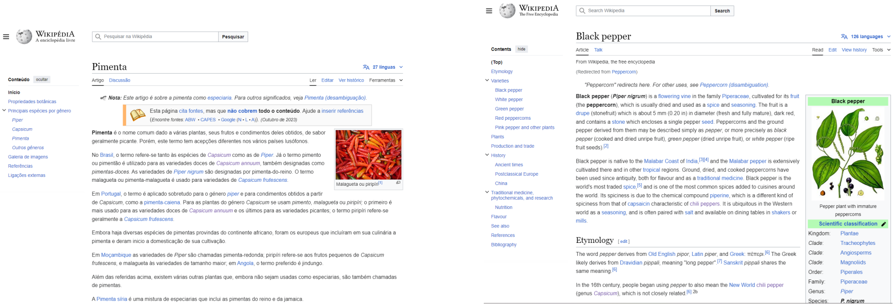
The accurate English equivalent of the Brazilian concept 'Pimenta' is 'chili pepper.' However, in Agrovoc [215] and Agrotermos [216], the concept of 'chili pepper' is incorrectly matched with the Portuguese term 'Condimento picante', which translates to 'spicy condiment' in English. This translation conflates two distinct concepts from different hierarchical levels, treating them as semantic equivalents when they are not.

This issue is not limited to Agrotermos and AGROVOC. On Wikipedia, a search for 'Pimenta' in Portuguese retrieves an article describing 'Pimentas' derived from *Capsicum* sp., as shown in Fig. 7.4a. However, switching the article's language to English leads to an article about 'pepper,' which refers to *Piper nigrum*, as illustrated in Fig. 7.4b.

This translation is problematic because the term 'Pimentas chili' is not even used in Brazil, leading to further confusion. These examples highlight how a single incorrect association can trigger a series of ontological inconsistencies across different terminology systems and languages.

To address this issue in APTO, we created two distinct classes, each explicitly linked to the correct biological classification:

- `Pepper: 'member of taxon' exactly 1` *`Piper nigrum`*



Pimenta

Nota: Este artigo é sobre a pimenta como especiaria. Para outros significados, veja Pimenta (desambiguação).

Esta página cita fontes, mas que **não cobrem todo o conteúdo**. Ajude a inserir referências (*Encontre fontes:* ABW • CAPES • Google (N • L • A)). *(Outubro de 2023)*

Pimenta é o nome comum dado a várias plantas, seus frutos e condimentos deles obtidos, de sabor geralmente picante. Porém, este termo tem acepções diferentes nos vários países lusófonos.

No Brasil, o termo refere-se tanto às espécies de *Capsicum* como as de *Piper*. Já o termo pimento ou pimentão é utilizado para as variedades doces de *Capsicum annuum*, também designadas como *pimentas-doces*. As variedades de *Piper nigrum* são designadas por pimenta-do-reino. O termo malagueta ou pimenta-malagueta é usado para variedades de *Capsicum frutescens*.

Malagueta ou piripíri[1]

Em Portugal, o termo é aplicado sobretudo para o género *piper* e para condimentos obtidos a partir de *Capsicum*, como a pimenta-caiena. Para as plantas do género *Capsicum* se usam *pimento*, *malagueta* ou *piripíri;* o primeiro é mais usado para as variedades doces de *Capsicum annuum* e os últimos para as variedades picantes; o termo piripíri refere-se geralmente a *Capsicum frutescens*.

Embora haja diversas espécies de pimentas provindas do continente africano, foram os europeus que incluíram em sua culinária a pimenta e deram inicio a domesticação de sua cultivação.

Em Moçambique as variedades de *Piper* são chamadas pimenta-redonda; piripíri refere-se aos frutos pequenos de *Capsicum frutescens*, e malagueta às variedades de tamanho maior; em Angola, o termo preferido é jindungo.

Além das referidas acima, existem várias outras plantas que, embora não sejam usadas como especiarias, são também chamadas de pimentas.

A Pimenta síria é uma mistura de especiarias que inclui as pimentas do reino e da jamaica.

Black pepper

From Wikipedia, the free encyclopedia

(Redirected from Peppercorn)

"Peppercorn" redirects here. For other uses, see Peppercorn (disambiguation).

Black pepper (***Piper nigrum***) is a flowering vine in the family Piperaceae, cultivated for its fruit (the **peppercorn**), which is usually dried and used as a spice and seasoning. The fruit is a drupe (stonefruit) which is about 5 mm (0.20 in) in diameter (fresh and fully mature), dark red, and contains a stone which encloses a single pepper seed. Peppercorns and the ground pepper derived from them may be described simply as *pepper*, or more precisely as *black pepper* (cooked and dried unripe fruit), *green pepper* (dried unripe fruit), or *white pepper* (ripe fruit seeds).[2]

Black pepper is native to the Malabar Coast of India,[3][4] and the Malabar pepper is extensively cultivated there and in other tropical regions. Ground, dried, and cooked peppercorns have been used since antiquity, both for flavour and as a traditional medicine. Black pepper is the world's most traded spice,[5] and is one of the most common spices added to cuisines around the world. Its spiciness is due to the chemical compound piperine, which is a different kind of spiciness from that of capsaicin characteristic of chili peppers. It is ubiquitous in the Western world as a seasoning, and is often paired with salt and available on dining tables in shakers or mills.

Etymology [edit]

The word *pepper* derives from Old English *pipor*, Latin *piper*, and Greek: πέπερι.[6] The Greek likely derives from Dravidian *pippali*, meaning "long pepper".[7] Sanskrit *pippali* shares the same meaning.[6]

In the 16th century, people began using *pepper* to also mean the New World chili pepper (genus *Capsicum*), which is not closely related.[6]:2b

Black pepper

Pepper plant with immature peppercorns

Scientific classification

Kingdom:	Plantae
Clade:	Tracheophytes
Clade:	Angiosperms
Clade:	Magnoliids
Order:	Piperales
Family:	Piperaceae
Genus:	*Piper*
Species:	***P. nigrum***

a Pimenta on Wikipedia in Portuguese.

b Black Pepper as Direct Translation of 'Pimenta' on Wikipedia in English.

Figure 7.4 'Pimenta' and its Translation as Black Pepper in Wikipedia. Sources: https://pt.wikipedia.org/wiki/Pimenta, https://en.wikipedia.org/wiki/Black_pepper

- Chili pepper: 'member of taxon' some *Capsicum*

Multilingual vocabularies often face ontological challenges due to the complexity of modeling concepts across diverse regions and languages. However, such issues are not exclusive to global contexts: regionalisms can also create significant ontological challenges at the local level. A prominent example is `apto:Cheiro_verde`, a concept representing a commonly used culinary herb mixture in Brazil. The term, which literally translates to 'green smell' in English, is a staple in Brazilian cuisine, enhancing the flavor of soups, stews, and meats.

The composition of 'Cheiro verde' varies significantly across Brazil's regions, as illustrated in Fig. 7.5. In the Southeast, South, and Midwest, it typically consists of parsley and chives. In the Northeast, parsley is replaced by coriander; in the North, the mixture expands to include chives, coriander, and fitweed. This regional variation is particularly important for traditional Amazonian dishes such as 'Pato no tucupi' (duck in tucupi) and 'Tacacá' [217, 218].

In APTO, the variations of this concept have been modeled as shown in Listing 7.7. This approach allows even non-Portuguese speakers to understand the basic semantics of the concept despite its lack of a direct English translation. The class `Cheiro_verde` is defined as a combination of `Coriander and Chives`, `Parsley and Chives`, or `Coriander and Chives and Fitweed`, with each ingredient linked to its corresponding species. These combinations should ideally be associated with the specific Brazilian regions to refine this modeling further.

Listing 7.7 'Cheiro verde' Representation in APTO

```
Class: 'Cheiro verde'
   SubClassOf: 'Culinary herbs'
```

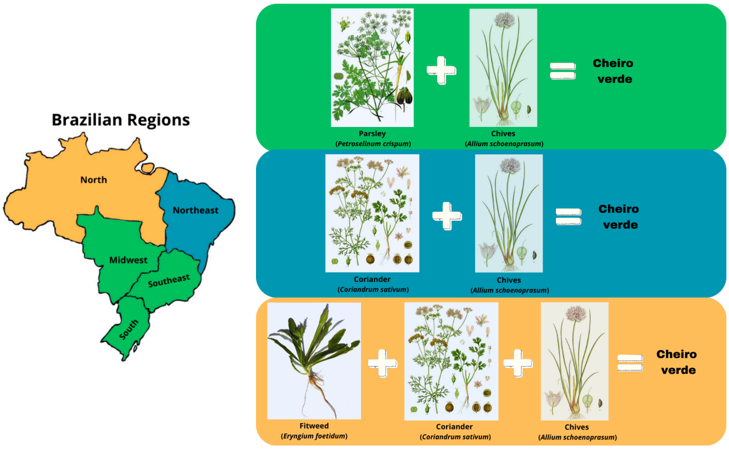


Figure 7.5 Herbs Included in ‘cheiro verde’ by Brazilian Regions. Parsley image source: [219]; Coriander image source: [220]; Chives image source: [221]; Fitweed image source: [222]. All images reused in this figure are licensed under the public domain on Wikimedia Commons.

```
   'Equivalent to'(Coriander and Chives) or
   (Coriander and Chives and Fitweed) or
   (Parsley and Chives)
...
Class: 'Coriander'
   'member of taxon' exactly 1 'Coriandrum sativum'

Class: 'Chives'
   'member of taxon' exactly 1 'Allium schoenoprasum'

Class: 'Fitweed'
   'member of taxon' exactly 1 'Eryngium foetidum'

Class: 'Parsley'
   'member of taxon' exactly 1 'Petroselinum crispum'
```

7.6.1 Other Design Choices Differing from Agrotermos and Agrovoc

Beyond the previously discussed cases, there are additional instances where the modeling in APTO diverges from that in Agrotermos and Agrovoc. These differences are detailed below.

Buriti: In APTO, the definition of *Buriti* differs from those in Agrotermos and Agrovoc. In Agrotermos, *Buriti* is not considered a preferred concept, whereas it is in APTO. Agrotermos and Agrovoc both favor *Mauritia*, a genus of oil palms, as the preferred term for *Buriti*. However, APTO distinguishes between the fruit, the plant, and the species' taxonomic identification as different classes. Therefore, adopting the same classification as Agrotermos and Agrovoc would be inconsistent. Consequently, the concept from Agrotermos is mapped to `apto:Buriti` as `skos:closeMatch`, while the concept from Agrovoc is mapped as `skos:relatedMatch`.

Coffee: The concept of 'Café' (coffe, in English) in Agrotermos is somewhat ambiguous. It suggests that 'Café' is the preferred term for 'cafeeiro' (the coffee plant), while also implying that 'Café' derives from 'Grãos de café' (coffee beans). Additionally, 'Grãos de café' is linked to 'Café' through a `hasProduct` relationship. The Agrotermos concept likely refers to the Agrovoc concept C_1731, which describes coffee as a beverage. In our ontology, we use the Agrovoc concept for coffee beans C_28379 to designate 'Café.' We define 'Café em grãos' as the preferred `skos:prefLabel`, and 'Café' as an `skos:altLabel` (since 'Café' is commonly used in databases to refer to 'Café em grãos,' it is included in the ontology as an alternate label). We did not model 'Café' as a beverage, as the beverage itself is not relevant in the context of trading.

Cashew: In Agrovoc, cashew is classified under 'Nuts,' while in Agrotermos, it is a subclass of 'Tropical Fruit.' The classification in Agrotermos aligns more closely with APTO, as it differentiates between fruit and nut. In APTO, 'Caju Nut' and 'Cashew' (representing the fruit) are modeled as separate classes, with the former as a subclass of 'Nuts' and the latter as a subclass of 'Tropical Fruit.' This distinction is crucial because the cashew fruit holds significant economic value beyond just the nut; for instance, cashew fruit pulp is recognized as a commodity by CONAB [126].

Coconut: Coconut is the term used in Agrotermos for the coconut from *Cocos nucifera*, but since there are other types of coconut, we modeled one class apto:Coconut with these other types of coconut as subclasses (e.g., Coco-da-baía, coco de babaçu, buriti).

Product types dependency: In cases where a processed product or by-product was identified for inclusion in APTO, but the corresponding raw product from which they originate was not, the raw product was also added. This is because processed products and by-products are inherently derived from the raw product, establishing a dependency relationship. For example, 'cassava flour' was listed but cassava was not, then we also included cassava to reflect this dependency.

Vegetal fiber: In APTO, 'vegetal fiber' has been modeled as a processed product. However, it can be considered both a raw and a processed product depending on the context. In some cases, such as when it is directly sourced from plant materials, it is regarded as a raw product. However, vegetal fiber often undergoes processing, such as cleaning, refining, or mechanical treatment, to be used in various applications like textiles or composites, making it a processed product. For instance, studies have indicated that vegetal fibers like those derived from bamboo can be utilized as raw materials, which are then processed for specific applications such as textile manufacturing or composite fabrication [223]. Additionally, research on the utilization of plant fibers in products like viscose rayon or polymer composites further supports the classification of vegetal fiber as a processed product after undergoing necessary processing steps [224]. Thus, whether vegetal fiber is considered raw or processed depends on its stage in the supply chain and the intended end-use.

7.7 Final Considerations

This study explored two approaches for integrating taxonomic data into APTO, leveraging ChatGPT with the BrowserOP plugin (Approach 1) and a Python algorithm developed with ChatGPT's assistance (Approach 2). Each approach demonstrated specific strengths and weaknesses, highlighting the potential and challenges of integrating AI and automation in ontology development.

Approach 1 utilized ChatGPT for direct interaction with the GBIF API, demonstrating its capability to handle complex queries, retrieve taxonomic information, and verify synonym relationships. However, scalability limitations became evident as the dataset size increased, leading to longer processing times and issues such as non-existing URIs and incomplete tasks. The subsequent discontinuation of GPT plugins further limited the reproducibility of this approach.

In contrast, Approach 2 addressed these limitations by employing a Python algorithm to process more extensive lists of species names efficiently. This approach successfully handled a list of 74 plant species in just over 2.5 minutes while maintaining consistent class-subclass relationships and avoiding redundancy in taxonomic hierarchies. However, it was sensitive to typographical errors, requiring a pre-processing step with ChatGPT for name validation.

The study underscores the complementary strengths of generative AI and algorithmic methods in ontology development. ChatGPT proved effective for tasks such as synonym detection, taxonomic hierarchy generation, and OWL code generation. Its ability to learn from errors and refine subsequent outputs demonstrates its potential for interactive and adaptive ontology design. In contrast, the Python algorithm provided the scalability and robustness required for handling larger datasets, demonstrating the practicality of combining AI-driven tools with structured programming techniques.

Challenges remain, particularly regarding dependency on external tools like ChatGPT, which may evolve unpredictably, as seen with the discontinuation of GPT plugins. Additionally, addressing semantic variations, such as those exemplified by the concepts ‘Pimenta’ and ‘Cheiro Verde,’ requires contextual modeling that goes beyond simple automation.

The workflow developed in this study is adaptable beyond the agricultural domain and can be applied to fields such as biodiversity informatics, ecological modeling, and conservation planning. Future work should explore fully automated solutions for error detection and validation, potentially incorporating machine learning models to reduce reliance on manual intervention. A complementary algorithm should be developed to periodically verify all species names in the ontology and update them in alignment with the latest version of the GBIF Backbone Taxonomy, for each new ontology version release.

In conclusion, this chapter demonstrates how combining AI-driven tools can streamline ontology engineering, address domain-specific challenges, and enhance scalability and efficiency.

CHAPTER 8

Ontology Testing and Validation

Chapter Overview

This chapter presents a novel approach to validating APTO by integrating both user UX metrics and FAIR assessment into the evaluation process. Traditional ontology validation often relies on verifying competency questions through SPARQL queries, yet it frequently overlooks the role of domain specialists in assessing usability and conceptual alignment. To bridge this gap, we employed Tree Testing to evaluate APTO, capturing specialists' navigation paths while solving 11 domain-specific tasks. This method, described in Section 8.4, provided actionable insights into the ontology's usability, highlighting structural weaknesses and areas for refinement. In addition to usability evaluation, we conducted a FAIR assessment to measure APTO's compliance with Findability, Accessibility, Interoperability, and Reusability principles, as described in Section 8.9. This process helped identify missing metadata, provenance details, and ontology relationships, allowing us to enhance its overall quality. By combining UX-driven methodologies with FAIR assessment, this study underscores the importance of a holistic approach to ontology validation. The findings demonstrate how usability testing and metadata refinement can work together to create more user-friendly and interoperable ontologies, fostering better knowledge representation for real-world applications.

Publication note: This chapter is based on the publication of Soares et al. [225]: *A Novel UX-Based Approach for Ontology Evaluation: Applying Tree Testing to the Agricultural Product Types Ontology.*

8.1 Introduction

As outlined in the ontology requirements elicitation in Section 6.4, we should conduct a usability evaluation of APTO. Ontology testing, as defined in the SABiO methodology by Falbo [73], involves the "verification and validation of the behavior of the operational ontology on a finite set of test cases, against the expected behavior regarding the competency questions" [73, p.8]. Essentially, this means that ontology testing is executed by implementing competency questions as queries in the chosen implementation environment.

However, the SABiO methodology provides a somewhat vague description of validation, particularly regarding the involvement of specialists in the validation process. While an ontology's ability to answer competency questions confirms its syntactical accuracy, it does not necessarily guarantee that the ontology accurately represents the domain or that domain specialists endorse the proposed model. This limitation is not unique to the SABiO methodology. As Casellas [226] notes, most ontology development methodologies offer only a vague description of the involvement of experts in ontology validation. Many ontologies undergo validation informally, often through undocumented discussions with specialists, which risks making the validation process opaque and potentially less reliable.

We argue that more rigorous, transparent, and informative methods are needed for testing and validating ontologies, particularly ontologies of types that function as classification systems. For these ontologies, it is critical that the proposed categorizations (hierarchies) accurately represent the domain and that this precision be validated by specialists. To achieve this, we validated APTO by using adapted UX metrics, specifically the Tree Testing protocol [227]. This approach offers a more comprehensive evaluation of the ontology's usability by allowing the creation of specific tasks and tracking how specialists (users) navigate through an ontology prototype to solve these tasks.

8.2 Related Work

Evaluating ontologies through UX metrics is not a widely adopted practice, but has been explored in some studies.

Casellas [226] pioneered the application of usability metrics to ontology evaluation, particularly within the legal domain. This study introduced the idea that usability measures, such as the System Usability Scale (SUS), can effectively assess the practical usability of ontologies from an end-user perspective. Casellas emphasized that while syntactical and structural correctness is important, the actual UX in interacting with the ontology is crucial. By considering how domain specialists and users perceive and use the ontology, this method provides a more holistic evaluation, ensuring that the ontology is functionally robust and user-friendly, which is an aspect often overlooked in traditional ontology validation processes.

Other research efforts have concentrated more on the usability of tools and techniques for ontology visualization [228, 229], as well as on the usability of ontology editors and engineering tools [230, 231], rather than directly ontology usability itself. This focus is likely due to the inherent complexity of evaluating ontology usability, particularly when using Task-Oriented Approaches like tree testing or SUS. As noted by Pak [232, p.12], “specifying the characteristics of ontologies is a complicated and time-consuming process; assessing its characteristics is quite subjective.” This subjectivity and complexity have made direct evaluations of ontology usability less common, despite their importance.

8.3 User Experience (UX)

The concept of UX emerged in the 1990s as a response to the growing complexity of digital systems and the need for a more holistic approach to understand how users interact with technology. The term “User Experience” was first popularized by Don Norman, a cognitive scientist and usability expert, during his time at Apple. Norman used the term to emphasize that design should take into account not only usability but also the overall experience of the user, including emotions, perceptions, and satisfaction [233]. This approach marked a shift from traditional usability, which focuses primarily on efficiency and effectiveness, towards a more comprehensive understanding of the user’s interaction with a product, service, or system [233].

8.3.1 Measuring User Experience

Although user experience is inherently complex and challenging to quantify, experts have developed a variety of techniques to measure UX in ways that yield actionable insights. These insights enable stakeholders to make informed decisions about improvements to products or interfaces. A seminal work in this area by Tullis and Albert [227] offers a robust framework for understanding different types of UX metrics and the contexts in which they should be applied.

Tullis and Albert [227] begin by differentiating between two primary types of study goals: formative and summative usability.

- **Formative Usability:** This approach is used during the early stages of design to identify potential issues and areas for improvement. It is iterative and diagnostic, aiming to inform the ongoing design process [227].
- **Summative Usability:** In contrast, summative usability is employed after the design is complete to evaluate its overall effectiveness. It is often comparative, used to benchmark against competitors or previous versions of the product [227].

The authors also introduce the concepts of performance and satisfaction as key user goals in UX evaluation.

- **Performance:** This metric focuses on how effectively users can complete tasks using the product. It includes factors such as task success rates, time on task, and error rates [227].
- **Satisfaction:** Satisfaction measures the subjective experience of users, typically assessed through self-reported metrics like surveys and questionnaires. It reflects the user's emotional response to the product [227].

8.3.2 Types of Usability Studies

Tullis and Albert [227] propose ten distinct types of usability studies, each serving a specific purpose:

1. Completing a transaction, which evaluates how easily users can accomplish a specific task, such as purchasing a product online.
2. Comparing products, where multiple products are assessed to determine which offers the best user experience.
3. Evaluating frequent use of the same product, focusing on how users interact with a product they use regularly.
4. Evaluating navigation or information architecture (IA), analyzing how easily users can find information within a system.
5. Increasing awareness, measures how effectively a product or feature informs or educates users.
6. Problem discovery, aimed at identifying issues that may hinder the user experience.
7. Maximizing usability for a critical product, ensures that essential products or features meet high usability standards.
8. Creating an overall positive user experience, addressing the general perception and emotional response of users.
9. Analyzing the impact of subtle changes, measuring how small modifications affect the user experience.
10. comparing alternative designs, which tests different design approaches to determine which is most effective.

8.3.3 Evaluation Methods

The authors categorize evaluation methods into three main types:

- **Traditional (Moderated) Usability Tests:** These involve direct observation of users as they interact with the product. Moderators can ask questions and probe deeper into user behaviors.
- **Online (Unmoderated) Usability Tests:** Conducted remotely, these tests allow users to complete tasks in their natural environment without real-time interaction with a moderator.

- **Online Surveys:** These gather self-reported data from users about their experiences and satisfaction levels.

8.3.4 Usability Metrics

Each of these methods can apply various usability metrics, which Tullis and Albert [227] group into the following categories:

- **Performance Metrics:** These include task success, time on task, error rates, efficiency, and learnability.
- **Issue-Based Metrics:** These focus on identifying specific usability issues encountered by users.
- **Self-Reported Metrics:** These encompass rating scales (such as Likert and Semantic Differential scales), post-task ratings, and post-session ratings, including tools like SUS (as in the case study of Casellas [226], described in the Section 8.2) and the Net Promoter Score (NPS).
- **Behavioral and Physiological Metrics:** These involve tracking user behavior, such as eye tracking (as in the case study of Fu [228], described in the Section 8.2), verbal expressions, and physiological responses like stress or emotional arousal.
- **Special Topics:** These specialized metrics combine different types of metrics, including web analytics (e.g., click-through rates and drop-off rates), A/B testing, card sorting, tree testing, and accessibility data.

8.4 APTO Evaluation Using Tree Testing

To evaluate APTO, we selected tree testing as the primary UX measuring protocol. We believe tree testing is particularly suitable for ontology evaluation. It combines various metrics, such as time on task and task success, to assess how users navigate and understand a hierarchy of concepts in an ontology prototype. In UX, a prototype is defined as a preliminary version of a product or system used to explore and evaluate usability. Prototypes can range from low-fidelity to high-fidelity, enabling designers and stakeholders to test ideas and gather feedback before finalizing the design [227].

APTO comprises a total of 581 classes, distributed across two modules: the Product Type module (288 classes) and the Organism module (293 classes). The present study focuses exclusively on the Product Type module, as it contains the conceptual hierarchies most relevant to agricultural commodity classification and user-centered evaluation. The Organism module was not included, since it only includes scientific names.

From the Product Type module, 204 classes were selected for inclusion in the prototype used in the tree testing. The remaining 84 classes were excluded because they directly corresponded to the task target concepts or contained parts of their names. Also, classes that were at the same level as the target concepts were also removed. Including these classes in the prototype would have compromised the validity of the test by revealing the correct answers.

For example, in Task 1, participants were asked to locate the most appropriate upper class for the concept *Whey*. If *Whey* had been included in the prototype, participants could simply navigate directly to it and complete the task without engaging in any meaningful evaluation of the taxonomy.

Usability testing with a prototype allows us to trace user paths through the ontology structure, providing insights into which aspects of the modeling may be unclear or inaccurate from the user's perspective. By analyzing user interactions, we can identify areas for improvement in the ontology design, enhancing its usability and ensuring it aligns more closely with users' domain models.

In the sequel, we discuss the configuration of the tree testing procedure in terms of its goals, participants, and methods.

8.4.1 Study Goals

From our perspective as the ontology designers, the usability test has been formative. Our primary goal was to detect potential issues and make improvements to a prototype of the ontology before its official release. This objective can be translated into key questions such as: *Can users effectively locate the concepts within the specified ontology hierarchy? What areas require improvement? What insights can we gain from analyzing users' navigation paths within the ontology?* To achieve this goal, we identified "areas of attention" in APTO, which were classified into two types: changes in the original hierarchy of classes imported from *Agrotermos* and *AGROVOC*, and new concepts (not existing in *Agrotermos* or *AGROVOC*) added to the APTO namespace.

From the users' perspective, the usability study was classified as performance-oriented, focusing on evaluating task success rates, time taken to complete tasks, and error rates. Each task involved working with one or more ontology fragments at a time. By ontology fragments, we refer to branches of the ontology tree, spanning from a broad concept to its most specific sub-concepts. Each task involved navigating one or more fragments to find the superclass to the target concept(s) on that task. These performance metrics offer valuable insights, including:

- **Time on Task:** Longer task completion times may suggest that users perceive the ontology fragments involved in the task as complex or confusing, highlighting areas that could benefit from simplification or reorganization. Conversely, shorter task times may indicate that the ontology fragments are well-structured and easy for users to navigate.

- **Task Success Rates:** High task success rates indicate that the ontology fragments used in a task effectively support users in finding information, validating the logical and intuitive arrangement of categories and relationships. In contrast, low success rates highlight problematic ontology fragments that may require refinement or reorganization.
- **Error Rates:** High error rates suggest that users often misunderstand the categorizations within the ontology fragments used in the task, highlighting a need for restructuring to enhance clarity and usability.

8.4.2 Participants

We refer to the participants in this study as “specialist users.” While the term is not formally defined in UX literature, it aligns the notion of a “power user” — individuals who utilize advanced features of a system more extensively and effectively than the average user [234]. In the context of our study, specialist users are defined as individuals who possessed two core areas of expertise:

- Experience in ontology engineering.
- Knowledge in the field of Brazilian agriculture, as the ontology is designed for potential users who are specialists working with product types within the Brazilian trading market.

Regarding the number of participants needed for usability testing, Nielsen [235] argued that testing with five users is often sufficient to uncover the majority of usability issues. Norman [233] supports this perspective and emphasizes that a small group of users can reveal the most critical problems, enabling rapid and cost-effective iterations. Tullis and Albert [227] also acknowledge the validity of this approach in many contexts. However, they caution that five users may not always be sufficient, particularly in cases where the user population is highly diverse or where the system being tested is complex and multifaceted. In such scenarios, they recommend increasing the sample size to ensure that the results are representative and that subtler issues, which might only be detected by a broader range of users, are also identified. However, in our case, the diversity of the target population is very limited, since ontology experts working in Brazilian agriculture represent a highly specialized and relatively small community. As such, the recruitment pool is constrained by the low availability of professionals who possess both technical ontology expertise and domain-specific knowledge in agriculture.

Initially, seven participants were recruited for the study. However, the results from one participant were excluded after their task completion time was identified as a statistical outlier (see Section 8.5). The final analysis therefore included data from six valid participants. Given the exploratory nature of the prototype being tested and the specificity of the intended user base, this number is aligned with established UX research guidelines for expert usability studies and was deemed sufficient to yield meaningful insights into the ontology's usability.

8.4.3 Usability Study Type and Evaluation Method

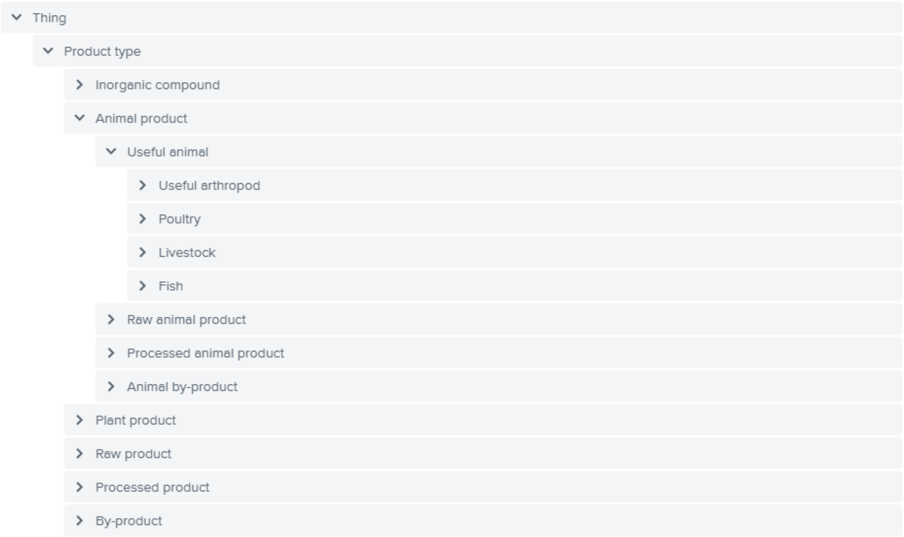
Our usability study focused on *Evaluating Navigation and/or Information Architecture*, as participants were tasked with navigating the ontology taxonomy to complete specific tasks. For the evaluation method, we selected an *Online (Unmoderated) Usability Test*, conducted remotely using the TreeJack tool from Optimal Workshop[1]. This study assessed an interactive prototype of the APTO Product Type module, allowing participants to interact with ontology classes to complete the assigned tasks. The prototype was presented as a navigation tree, as illustrated in Fig. 8.1. The full taxonomy of the Product Type module is available in the Data Availability Section (Dataset 14.).

This module is structured as a polyhierarchy, organized into multiple hierarchical levels based on two classification dimensions: origin (including categories like *Inorganic Compound*, *Animal-Origin Product*, and *Plant-Origin Product*), and type of processing (including categories like *Raw Product*, *Processed Product*, and *By-product*), shown in Fig. 8.1. This means that tasks could have multiple correct answers within the ontology tree. A brief description was included in the instructions to clarify the meaning of each of these upper classes for the participants.

Eleven tasks were designed, each focused on a specific target concept or group of target concepts. Participants were required to navigate the prototype to identify the most appropriate superclass in the ontology for the target concept(s). The exact names of the target concepts were not visible in the navigation tree, as recommended for this type of protocol [227], requiring participants to explore the categories and make decisions based on the provided descriptions. This setup simulated real-world scenarios where users might need to classify items with limited information. Although the polyhierarchical nature of the ontology allowed for a target concept to belong to more than one superclass, participants were instructed to select only one class as the most appropriate answer for each task. This should reveal navigation preferences within the ontology.

---

[1] `https://www.optimalworkshop.com/product/tree-testing`



- Thing
  - Product type
    - Inorganic compound
    - Animal product
      - Useful animal
        - Useful arthropod
        - Poultry
        - Livestock
        - Fish
      - Raw animal product
      - Processed animal product
      - Animal by-product
    - Plant product
    - Raw product
    - Processed product
    - By-product

Figure 8.1 APTO Prototype Upper Classes.

The prototype, instructions, and all supplementary materials used in this study were originally provided to participants in Portuguese and have been translated into English to be included in this dissertation.

8.4.4 Pre-Test

A pre-test was conducted with two participants to identify and resolve potential issues in this study protocol. Initially, the study included 12 tasks, but two tasks were found to be very similar and had consistency issues, as noted by one of the pre-test participants. As a result, one task was eliminated from the study. Additionally, the pre-test revealed an oversight where certain correct paths in the prototype were not properly configured, leading to false “wrong answers” being recorded. These issues were corrected prior to the commencement of the study. Pre-test participants did not participate in the actual study.

8.5 Results

Table 8.1 displays the overall results by participants, including the outlier (participant 3), who was excluded from the analysis. Notably, no task was skipped. Detailed average scores for each task are presented and discussed in Figures 8.2 to 8.12[2].

---

[2] Although the tasks are listed sequentially here, participants received them in varying orders, in accordance with the randomization function of TreeJack.

To identify the outlier in task completion times, we used the interquartile range (IQR) measure. Let T be the set of completion times in minutes. The lower quartile Q_1 and upper quartile Q_3 are defined as:

$$Q_1 = 25\text{th percentile of } T, \quad Q_3 = 75\text{th percentile of } T$$

The interquartile range is:

$$\text{IQR} = Q_3 - Q_1$$

An outlier is any time $t \in T$ such that:

$$t > Q_3 + 1.5 \times \text{IQR}$$

Using the observed values:

$$Q_1 = 11.62, \quad Q_3 = 16.19, \quad \text{IQR} = 4.57$$

$$\text{Outlier threshold} = 16.19 + 1.5 \times 4.57 = 23.05 \text{ minutes}$$

Any completion time greater than 23.05 minutes was therefore considered a statistical outlier and excluded from the analysis.

Table 8.1 Participant Task Performance Summary

Participant	Status	Time Taken	Decimal Min.	Question Responses	Tasks Completed	Tasks Skipped	Tasks Successful
1	Compl.	00:18:41	18.68	2	100%	0%	54%
2	Compl.	00:11:35	11.58	2	100%	0%	27%
3	Compl.	06:23:49	383.82	2	100%	0%	45%
4	Compl.	00:08:49	8.82	2	100%	0%	36%
5	Compl.	00:12:44	12.73	2	100%	0%	72%
6	Compl.	00:13:42	13.70	2	100%	0%	90%
7	Compl.	00:11:39	11.65	2	100%	0%	63%

In each chart from Figures 8.2 to 8.12, the following scores are included [236]:

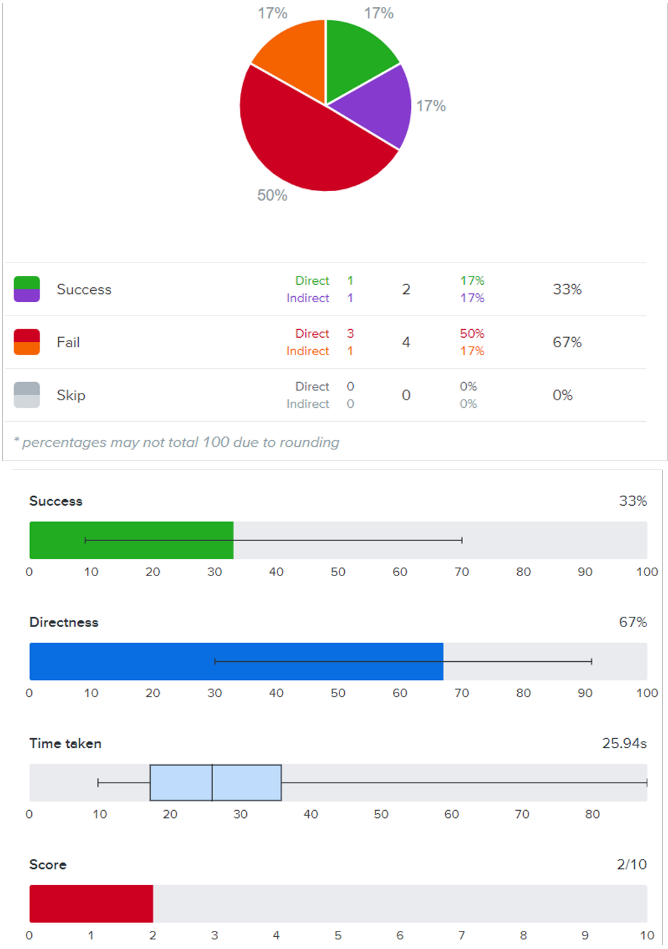
- **Success Score:** This represents the percentage of participants who successfully navigated to the correct concept for the task. For each task, there is one or multiple intended correct paths in the ontology.
  - **Direct Success:** Participants who navigated directly to the correct superclass without deviation.

  - **Indirect Success:** Participants who initially navigated down the wrong path, backtracked, and then successfully found the correct superclass.
- **Failure Score:** This measures the percentage of participants who chose an incorrect superclass.
  - **Direct Fail:** Participants who navigated directly to an incorrect superclass without backtracking.
  - **Indirect Fail:** Participants who initially navigated down the wrong path, backtracked, and still ended up at an incorrect superclass.
- **Skip Score:** The percentage of participants who skipped the task before selecting any concept.
  - **Direct Skip:** Participants who clicked the "skip" button without interacting with the tree.
  - **Indirect Skip:** Participants who started navigating through the tree but then chose to skip the task.
- **Directness Score:** This indicates the percentage of participants who took a direct path toward their selected answer without backtracking. This score is particularly useful when compared to the success score to determine whether participants were truly successful or if they had to correct their path mid-way.
- **Time Taken:** The average time, in seconds, that participants took to complete the task. This median time is illustrated by the line in the middle of the light blue box on the charts.
- **Overall Score:** A weighted average of the success and directness scores for each task. An overall score of 7 or higher generally indicates good performance.
- **Abandoned:** This score represents instances where a task was neither completed nor skipped, often due to the participant closing the session or timing out.

In the sequel, we first present the individual results for each task and subsequently provide a comparative analysis in Section 8.6. A more detailed analysis of these results is provided in Section 8.7.

8.5.1 Task 1 - Target Concept: *Whey*

The goal of this task was to determine whether the specialists would categorize *whey* as a by-product, or more specifically, as a *milk by-product*. The interest for this concept validation arises because the classification proposed in APTO differs from that in AGROVOC: while AGROVOC categorizes whey as a processed milk product, APTO classifies it under *milk by-product*. Fig. 8.2 shows that only two participants selected one of the two correct paths, resulting in a 33% success rate. The directness score of 67% indicates that most participants were confident in their choices, as they did not backtrack. The average time to complete this task was 25.94 seconds. The overall task score was 2, reflecting poor performance on this task.



🟩 🟪	Success	Direct Indirect	1 1	2	17% 17%	33%
🟥 🟧	Fail	Direct Indirect	3 1	4	50% 17%	67%
⬜ ⬜	Skip	Direct Indirect	0 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.2 Task 1 - **Question:** 'Whey' is the liquid remaining after milk coagulates during cheese production. Rich in proteins, vitamins, and minerals, whey is widely used in the food industry, especially in nutritional supplements, bakery products, and beverages. Which class in the ontology do you consider the most suitable as the upper class for 'whey'?
Correct Paths:
1) Thing > Product type > Animal product > Animal by-product > Milk by-product
2) Thing > Product type > By-product > Animal by-product > Milk by-product

8.5.2 Task 2 - Target Concepts: *Soybean Meal* and *Wheat Bran*

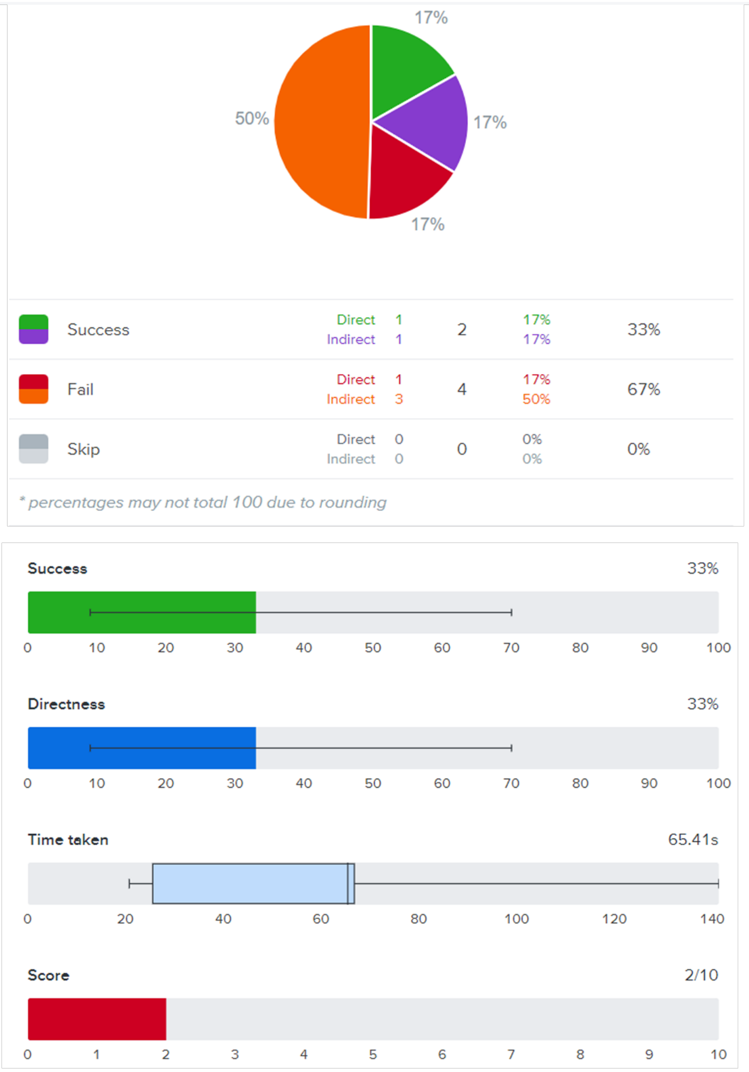
The objective of Task 2 was to determine whether the participants would categorize *Soybean meal* and *Wheat bran* as *Plant by-products*. This task aimed to validate the modeling of these two new concepts, which were added to the APTO namespace because they were not found in AGROVOC or Agrotermos. Fig. 8.3 shows that only two participants selected one of the two correct paths, resulting in a 33% success rate. However, the directness score for this task was 33%, indicating that participants frequently backtracked during the task. This backtracking directly influenced the average time to complete the task, which was 65.41 seconds. The overall task score was 2, reflecting poor performance.

8.5.3 Task 3 - Target Concepts: *'Boi Magro'*, *'Boi Gordo'*, *'Vaca Gorda'*, and *'Vaca Leiteira'*

'Boi Magro', *'Boi Gordo'*, *'Vaca Gorda'*, and *'Vaca Leiteira'* are terms used in Brazilian trade to indicate different categories of bovine, such as an ox or cow ready for slaughter (*'Boi Gordo'* and *'Vaca Gorda'*), an ox that still needs to be fattened before slaughter (*'Boi Magro'*), and a cow designated for milk production (*'Vaca Leiteira'*). These concepts are specific to the trading market in Brazil and lack direct translations into English. Additionally, they are not included in Agrotermos or AGROVOC, which are broader and more generalist vocabularies. Therefore, we incorporated these new concepts into the APTO namespace under the upper class *Bovine* (see the complete hierarchy in Fig. 8.4). This task aimed to validate the modeling of these new concepts within the *Product Types* module. Fig. 8.4 shows that four participants selected the correct path, resulting in a 67% success rate. The directness score was 50%, indicating that participants moderately backtracked. The average time to complete the task was 35.58 seconds. The overall task score was 4, well below the ideal score of 7.

8.5.4 Task 4 - Target Concepts: *Frozen Chicken* and *Chilled Chicken*

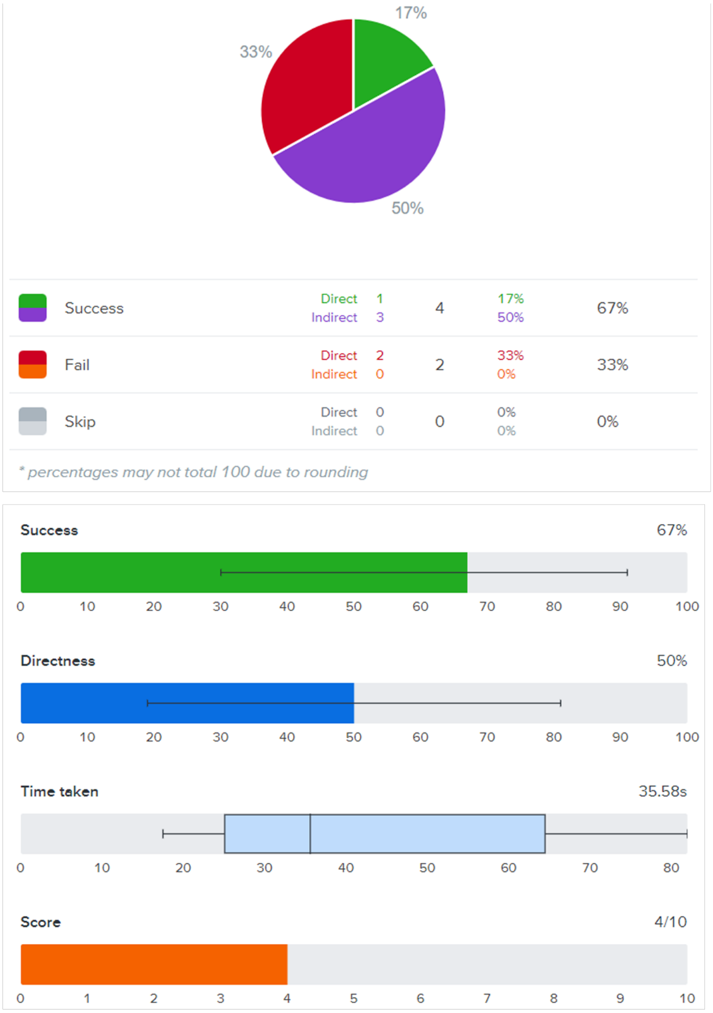
Frozen chicken and *Chilled chicken* refer to chicken meat *in natura*, which undergoes different temperatures of refrigeration to prevent spoilage [237]. Although these meats undergo basic processing (slaughtering and cutting), this processing does not alter the product's composition, so it is still considered a raw product. These two concepts were not found eighter in the Agrotermos nor AGROVOC namespaces; therefore, we added them to the APTO namespace. The aim of this task was to validate the modeling of these new concepts. As shown in Fig. 8.5, four participants selected one of the four correct paths, resulting in a 67%



🟩	Success	Direct Indirect	1 1	2	17% 17%	33%
🟥	Fail	Direct Indirect	1 3	4	17% 50%	67%
▨	Skip	Direct Indirect	0 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.3 Task 2 - **Question:** 'Soybean meal' and 'Wheat bran' are obtained from the processing of soybeans and wheat during the production of other primary items, but they are not the intended final product. If you had to choose a single class to group the terms 'Soybean meal' and 'Wheat bran', which class in the ontology would you choose?
Correct Paths:
1) Thing > Product type > Plant product > Plant by-product
2) Thing > Product type > By-product > Plant by-product

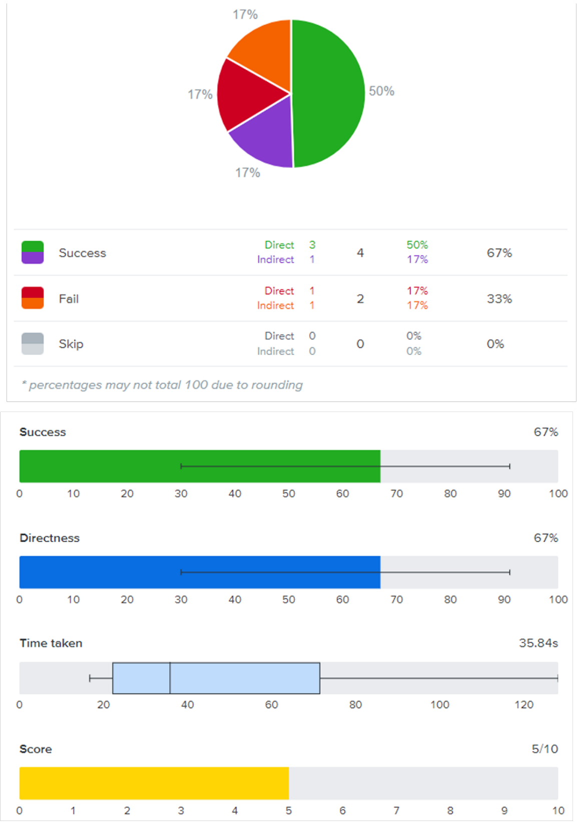


🟩🟪	Success	Direct 1 Indirect 3	4	17% 50%	67%
🟥🟧	Fail	Direct 2 Indirect 0	2	33% 0%	33%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.4 Task 3 - **Question:** 'Boi magro' and 'boi gordo' are terms commonly used in the financial market to differentiate between adult cattle ready for slaughter (boi gordo) and those that still need care to reach that point (boi magro). Similarly, 'vaca gorda' refers to a cow ready for slaughter, while 'vaca leiteira' describes a cow intended for milk production. In the ontology, which class do you believe could serve as the 'parent' class capable of grouping all these terms?
Correct Path:
1) Thing > Product type > Animal product > Useful animal > Livestock > Bovine

success rate. The directness score was 67%, indicating that participants had less difficulty navigating the ontology for this task. The average time to complete the task was 35.84 seconds. The overall task score was 5.



🟩🟪	Success	Direct 3 Indirect 1	4	50% 17%	67%
🟥🟧	Fail	Direct 1 Indirect 1	2	17% 17%	33%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.5 Task 4 - **Question:** 'Frozen chicken' and 'chilled chicken' are common terms used to describe slaughtered chicken that undergoes little or no processing beyond slaughter and cleaning and is sent directly from farms to distributors such as cold storage facilities and supermarkets. In the ontology, which class do you believe could serve as the upper class capable of grouping these two terms?
Correct Paths: 1) Thing > Product type > Animal product > Raw animal product > Livestock product > Poultry product > Chicken meat 2) Thing > Product type > Animal product > Raw animal product > Livestock product > Meat > Chicken meat 3) Thing > Product type > Raw product > Raw animal product > Livestock product > Poultry product > Chicken meat 4) Thing > Product type > Raw product > Raw animal product > Livestock product > Meat > Chicken meat

8.5.5 Task 5 - Target Concepts: *Dairy Drink*, *Yogurt*, *Powdered Milk*, *Pasteurized Milk*, *UHT Milk*, and *Cheese*

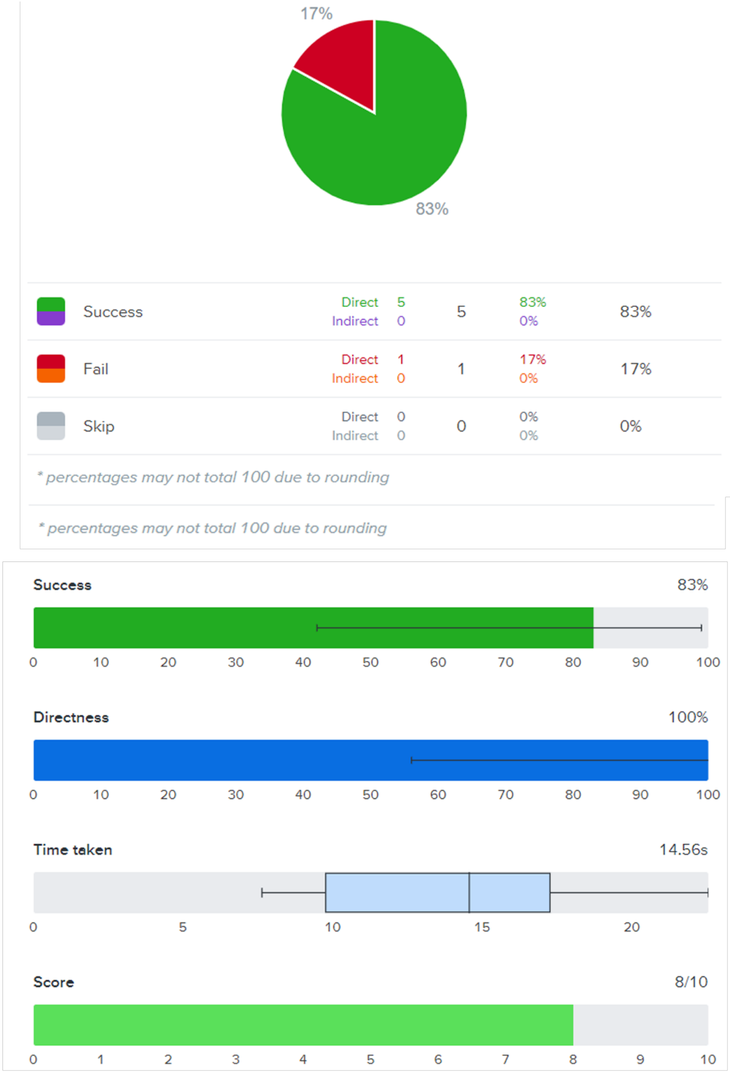
All the concepts in this task refer to processed milk products. With the exception of *Dairy Drink*, all other concepts were imported from AGROVOC. *Dairy Drink* is the closest translation of the Portuguese term *Bebida láctea*, which has no direct equivalent in English. As explained in Section 6.6.1, is a beverage mix of whey and milk, and is therefore considered a processed product. The aim of this task was to validate the structure incorporated from AGROVOC and the newly added concept. The success rate for this task was significantly higher than in the previous ones, at 83%, as shown in Fig. 8.6. Additionally, the directness score was 100%, indicating that participants encountered no difficulties navigating to the correct answer. The average time taken was 14.56 seconds, which was also much shorter than in previous tasks. The overall task score was 8, which is considered a good score (above 7) [236]. This suggests participants found this part of the ontology straightforward to navigate.

8.5.6 Task 6 - Target Concepts: *Araucaria*, *Eucalyptus*, and *Pinus*

These three types of tree were added to the APTO namespace and are not included in AGROVOC nor Agrotermos in the same way they are represented in APTO. They were categorized as useful plants in APTO, as they are sources of a myriad of products, such as cellulose and wood. The aim of this task was to validate the proposed categorization. The success rate for this task was 17%, as shown in Fig. 8.7, the lowest of all tasks in this study. This suggests a potentially critical error in the ontology modeling, which prevented users from finding the correct solution. This may be related to the fact that these are forestry products, which users may be less familiar with. However, the directness score for this task was high, at 83%, and the time on task was low, at 27.48 seconds, indicating that users did not get lost and were confident in the answers they provided. The overall task score was 2, which is considered poor.

8.5.7 Task 7 - Target Concepts: *Andiroba almond*, *Cocoa bean*, *Cashew nut*, and *Coconut*

These products are both edible nuts, as they can be consumed by humans, and oil seeds, as they are also used by the industry to produce oil. The aim of this task was to validate the addition of these new concepts to the APTO schema. Four correct paths were defined for this task, and as shown in Fig. 8.8, 100% of the participants successfully completed

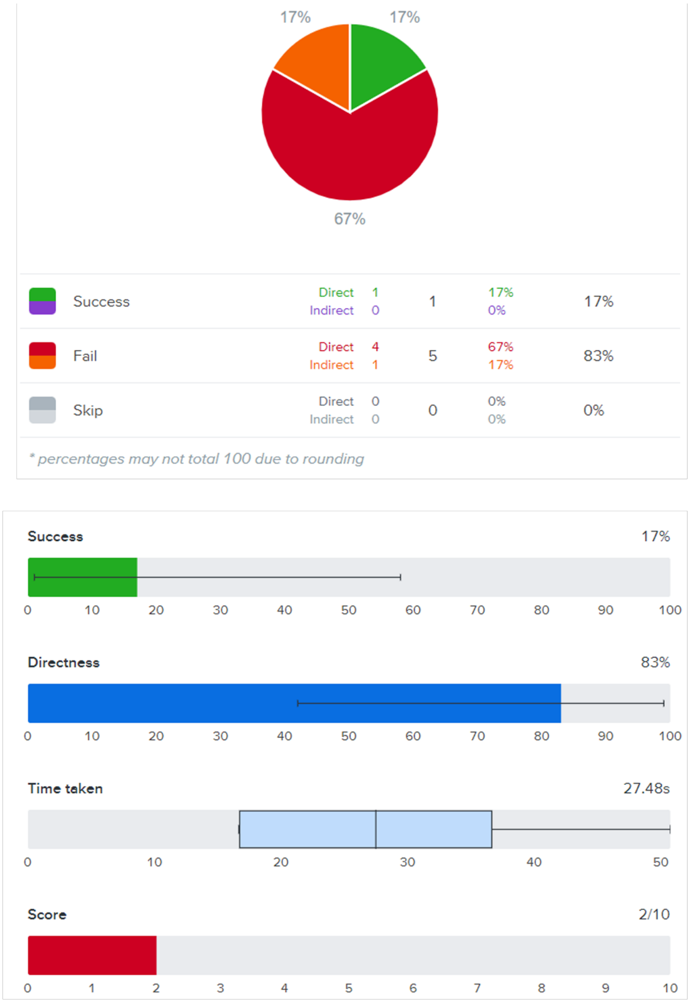


🟩	Success	Direct 5 Indirect 0	5	83% 0%	83%
🟥	Fail	Direct 1 Indirect 0	1	17% 0%	17%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

* percentages may not total 100 due to rounding

* percentages may not total 100 due to rounding

Figure 8.6 Task 5 - **Question:** ‘Dairy drink’, ‘yogurt’, ‘powdered milk’, ‘pasteurized milk’, ‘UHT milk’, and ‘cheese’ all have something in common: milk as the main ingredient. Which class in the ontology could be considered the ”parent” of these terms?
Correct Paths:
1) Thing > Product type > Animal product > Processed animal product > Milk product
2) Thing > Product type > Processed product > Processed animal product > Milk product



🟩	Success	Direct 1 Indirect 0	1	17% 0%	17%
🟥	Fail	Direct 4 Indirect 1	5	67% 17%	83%
▨	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.7 Task 6 -**Question:** 'Araucaria', 'eucalyptus', and 'pinus' are commercial names for trees of the species Araucaria sp., Eucalyptus sp., and Pinus sp., respectively. The cultivation of these species aims at the production of wood and its derivatives, such as cellulose, charcoal, etc. Which upper class in the ontology would best fit the terms araucaria, eucalyptus, and pinus?

Correct Paths:

1) Thing > Product type > Plant product > Useful plant > Cellulose-producing plant
2) Thing > Product type > Plant product > Useful plant > Wood-producing plant

the task. However, the directness score was low, at 33%, and the time on task was high, at 42.34 seconds, indicating that participants found it challenging to locate the correct answer and had to backtrack multiple times. Despite these challenges, the overall score was good, at 8.

8.5.8 Task 8 - Target Concepts: *Pineapple pulp*, *Açaí pulp*, *Buriti pulp*, *Ceriguela pulp*, and *Cupuaçu pulp*

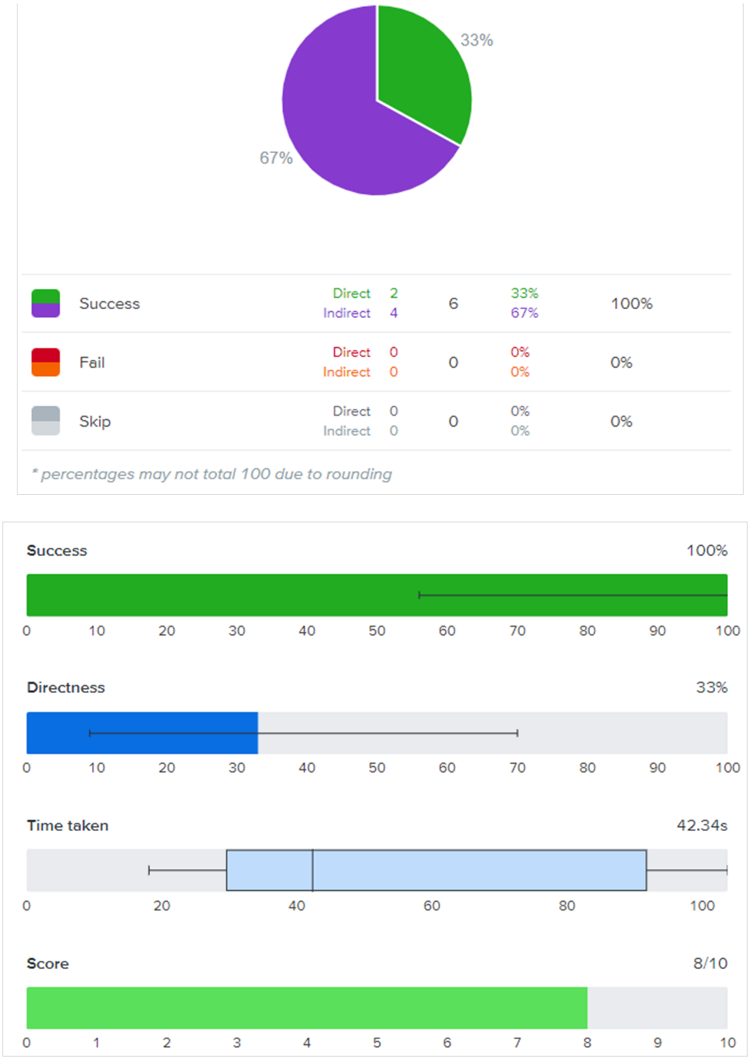
These concepts are also new additions to the APTO namespace and did not exist in AGROVOC nor in Agrotermos. When participants navigated the ontology to locate the parent class for these concepts, 67% successfully completed the task. The directness score was moderate, at 50%, indicating that some participants experienced difficulties in finding the correct answer. The time on task was 37.16 seconds, which further supports the hypothesis that participants faced challenges during this task. The overall score was low, at 4, potentially indicating issues in this ontology fragment.

8.5.9 Task 9 - Target Concepts: *Rice flakes* and *Corn flakes*

Rice flakes and *Corn flakes* are processed cereal products obtained from rice and corn, respectively. These concepts are also new additions to the APTO namespace. The success rate for this task was moderate, with only 50% of participants successfully completing it, as shown in Fig. 8.10. The directness score was also 50%, and the time on task was 29.13 seconds, suggesting that participants may have experienced difficulty navigating the ontology to find the correct paths. Consequently, the overall score was low, at 3.

8.5.10 Task 10 - Target Concepts: *Cane sugar* and *Ethanol*

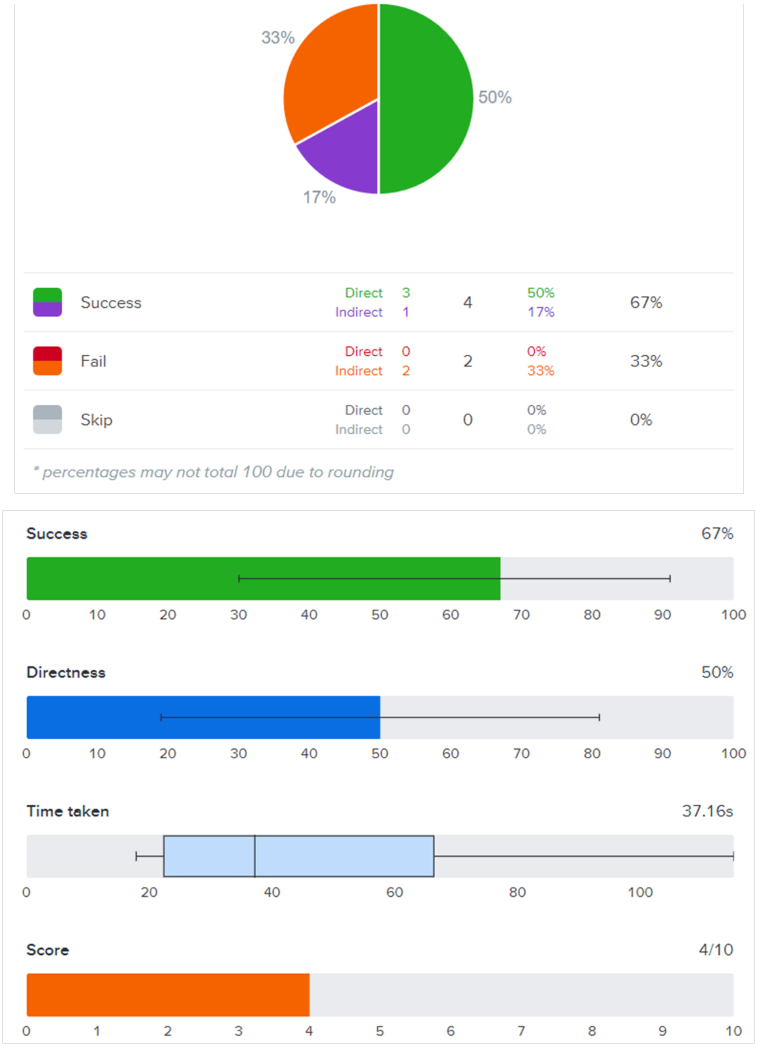
Cane sugar and *Ethanol* are two processed products derived from sugar cane. Although these concepts are not new and are included in the AGROVOC and Agrotermos namespaces, the upper class defined for them in APTO is new. The term ‘Sucroenergético’, which has no direct translation to English, is used in Brazil to group these two products derived from sugar cane. The aim of this task was to validate this new upper class. Only 33% of participants successfully completed this task, as shown in Fig. 8.11. However, the directness score was high (67%) and the time on task was relatively low (27.74 seconds), indicating that participants were confident in their answers. The resulting overall score was 2, suggesting a potential need for improvements in this ontology fragment.



🟩	Success	Direct Indirect	2 4	6	33% 67%	100%
🟥	Fail	Direct Indirect	0 0	0	0% 0%	0%
▨	Skip	Direct Indirect	0 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

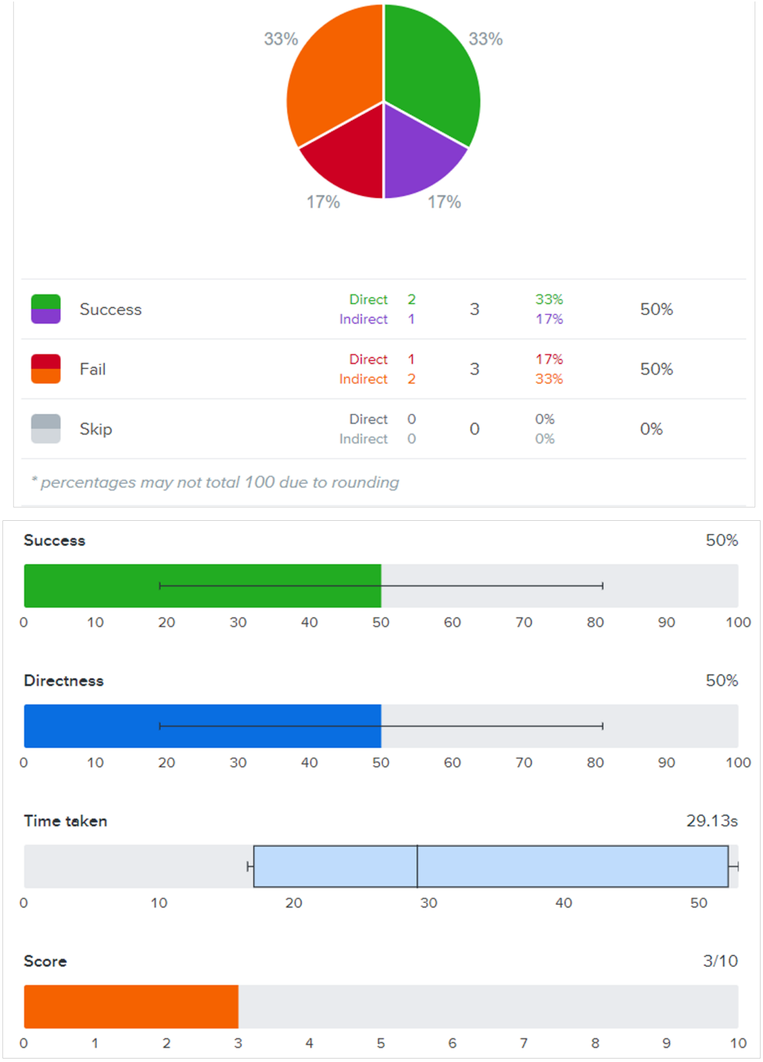
Figure 8.8 Task 7 - **Question:** ‘Andiroba almond’, ‘cocoa bean’, ‘cashew nut’, and ‘coconut’ are examples of oilseeds that can be used for human consumption. In which category of the ontology would you classify them?
Correct Paths:
1) Thing > Product type > Plant product > Raw plant product > Edible nut
2) Thing > Product type > Plant product > Raw plant product > Oil seed
3) Thing > Product type > Raw product > Raw plant product > Edible nut
4) Thing > Product type > Raw product > Raw plant product > Oil seed



🟩🟪	Success	Direct Indirect	3 1	4	50% 17%	67%
🟥🟧	Fail	Direct Indirect	0 2	2	0% 33%	33%
⬜	Skip	Direct Indirect	0 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

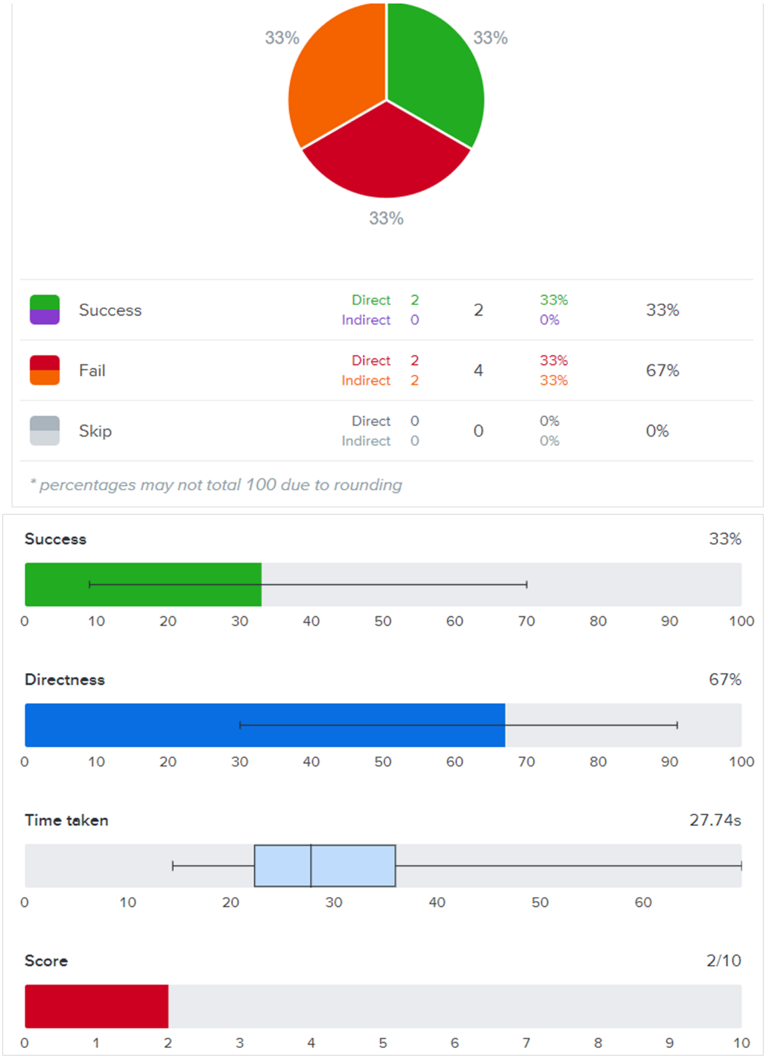
Figure 8.9 Task 8 - **Question:** ‘Pineapple pulp’, ‘açaí pulp’, ‘buriti pulp’, ‘ceriguela pulp’, and ‘cupuaçu pulp’ are examples of fruit pulps marketed in Brazil. What would be the most suitable upper class in the ontology to group these terms?
Correct Paths:
1) Thing > Product type > Plant product > Processed plant product > Fruit pulp
2) Thing > Product type > Processed product > Processed plant product > Fruit pulp



🟩🟪	Success	Direct 2 Indirect 1	3	33% 17%	50%
🟥🟧	Fail	Direct 1 Indirect 2	3	17% 33%	50%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.10 Task 9 - **Question:** ‘Rice flakes’ and ‘corn flakes’ are examples of products processed from cereals. Which class in the ontology could be considered the most suitable parent class for these products?
Correct Paths:
1) Thing > Product type > Plant product > Processed plant product > Cereal product
2) Thing > Product type > Processed product > Processed plant product > Cereal product



🟩🟪	Success	Direct 2 Indirect 0	2	33% 0%	33%
🟥🟧	Fail	Direct 2 Indirect 2	4	33% 33%	67%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.11 Task 10 - **Question:** ‘Cane sugar’ and ‘ethanol’ are examples of products derived from sugarcane. Which class in the ontology do you consider the most suitable upper class for these products?
Correct Paths:
1) Thing > Product type > Plant product > Processed plant product > Sucroenergético
2) Thing > Product type > Processed product > Processed plant product > Sucroenergético

8.5.11 Task 11 - Target Concepts: *Babassu oil*, *Macauba oil*, *Buriti oil*, *Copaiba oil*, *Murumuru oil*, and *Pequi oil*

These concepts represent different types of oil seeds that are widely used in oil production in Brazil. In the prototype, there were two correct paths, reflecting the APTO categorization of these classes. These classes do not exist in the Agrotermos or AGROVOC namespaces, so the aim of this task was to validate these new classes added to the APTO namespace. The success rate for this task was considerably high, at 83%, with the same score for directness, as shown in Fig. 8.12. The average time taken was also low, at 23.54 seconds. The overall score was 7, which is considered good.

8.6 Comparative Analysis of Task Results

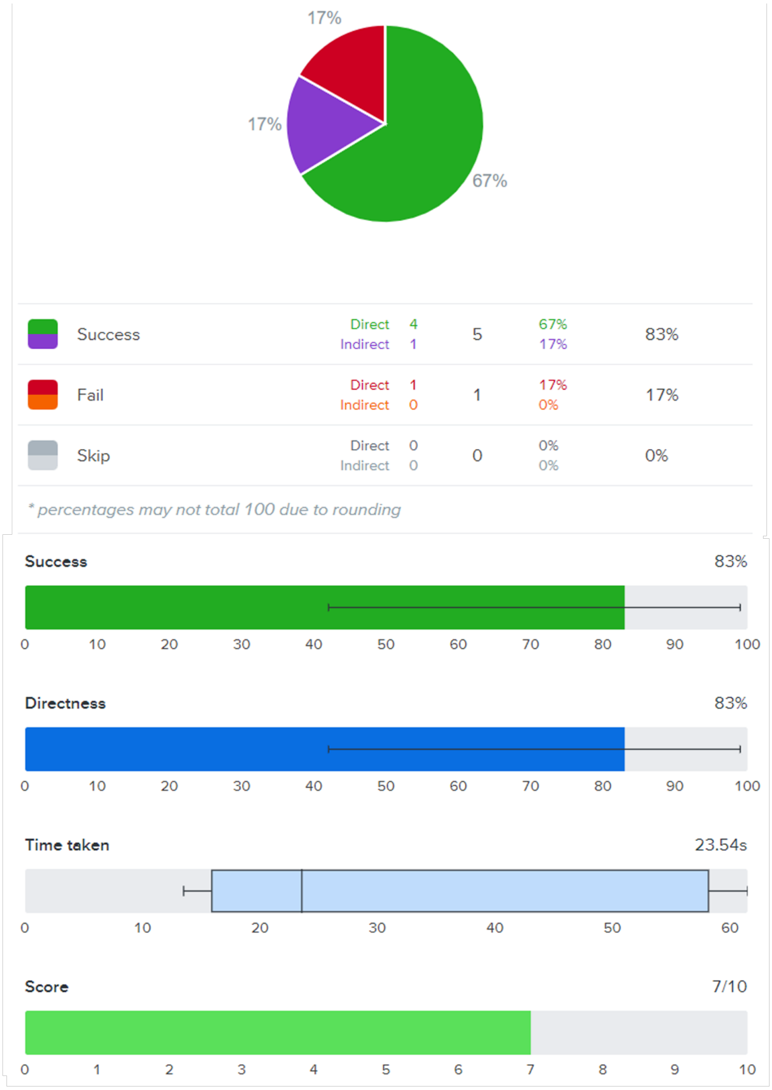
Based on the results from all 11 tasks (also shown in Table 8.2), the comparative analysis reveals clear trends, areas of strength, and potential weaknesses in the ontology design, as reflected by participant performance:

Table 8.2 Task Performance Summary

Task	Success (%)	Directness (%)	Time Taken (s)	Score
Task 1	33	67	25.94	2
Task 2	33	33	65.41	2
Task 3	67	50	35.58	4
Task 4	67	67	35.84	5
Task 5	83	100	14.56	8
Task 6	17	83	27.48	2
Task 7	100	33	42.34	8
Task 8	67	50	37.16	4
Task 9	50	50	29.13	3
Task 10	33	67	27.74	2
Task 11	83	83	23.54	7

Success and Directness

- **High Success and Directness**: Tasks 5 and 11 stand out with success rates of 83%, high directness scores (100% and 83%, respectively), and relatively low completion times. These results indicate that participants found these fragments of the ontology intuitive and easy to navigate, reflecting effective modeling.
- **Moderate Success with Low Directness**: Tasks such as 3, 4, 7, and 8 had success rates of 67% but varied in directness. For example, Task 7 achieved 100% success, but the directness score was only 33%, suggesting that participants struggled with initial navigation and had to backtrack before arriving at the correct solution.



🟩🟪	Success	Direct 4 Indirect 1	5	67% 17%	83%
🟥🟧	Fail	Direct 1 Indirect 0	1	17% 0%	17%
⬜	Skip	Direct 0 Indirect 0	0	0% 0%	0%

** percentages may not total 100 due to rounding*

Figure 8.12 Task 11 - **Question:** ‘Babassu oil’, ‘macaúba oil’, ‘buriti oil’, ‘copaiba oil’, ‘murmuru oil’, and ‘pequi oil’ are examples of vegetable oils extracted from different fruits and seeds native to Brazil. Which class in the ontology do you consider the most suitable ”parent” class for these products?
Correct Paths:
1) Thing > Product type > Plant product > Processed plant product > Plant oil
2) Thing > Product type > Processed product > Processed plant product > Plant oil

- **Low Success Rates**: Tasks 1, 2, 6, and 10 had success rates of 33% or lower. Task 6 was particularly notable for its poor performance (17% success), despite a high directness score of 83%. This discrepancy suggests that participants confidently chose incorrect answers, highlighting potential issues in the ontology's structure or categorization logic.

Time on Task

- **Quick Completion**: Tasks 5 and 11, with low average times of 14.56 seconds and 23.54 seconds, respectively, indicate clear and straightforward navigation for participants.
- **High Completion Times**: Tasks 2 and 7 had significantly higher times (65.41 and 42.34 seconds, respectively). These longer durations align with their lower directness scores, indicating that participants spent additional time navigating or backtracking.

Overall Performance

- **Best Performance**: Task 5 achieved the highest overall score (8), with perfect directness and high success rates. Similarly, Task 7 also scored 8 due to its perfect success rate, despite the lower directness score.
- **Worst Performance**: Tasks 1, 2, 6, and 10 had overall scores of 2, reflecting significant usability challenges in those fragments of the ontology. Task 6, in particular, indicates a critical need for redesign due to the very low success rate.

Trends and Insights

- **Concept Familiarity**: Tasks involving more familiar or well-established concepts (e.g., processed milk products in Task 5) tended to perform better. In contrast, tasks with less familiar or newly introduced concepts (e.g., tree categorization in Task 6) struggled, suggesting the need for clearer definitions or improved categorization.
- **Ontology Structure and Relationships**: Tasks 7 and 10 highlight the impact of ambiguous or complex relationships between classes. For example, backtracking in Task 7 suggests that participants struggled to distinguish between overlapping categories (e.g., nuts as oil seeds vs. edible products).
- **Task Complexity**: Tasks with multiple correct paths (e.g., Task 11) tended to perform well, as participants had more flexibility in identifying valid answers. However, tasks with stricter paths (e.g., Task 2) presented greater challenges, potentially requiring adjustments to the ontology or the task design.

8.7 Detailed Analysis of Results

The charts presented in the Section 8.5 are valuable for understanding the overall performance of participants in each task; however, they provide limited insight into the specific paths participants most frequently chose and the reasons behind their choices. Therefore, in this section, we introduce the so-called Pietrees, which allow us to delve deeper into the details and better comprehend the outcomes for each task. Analyzing these results enabled us to make improvements to APTO and discuss the participants' choices for each task.

8.7.1 Interpreting Pietrees: Insights into User Navigation

A Pietree is composed of pies (nodes) representing the concepts in the information architecture (classes in our ontology) and lines depicting the users' navigation paths. The size of a node and the thickness of a branch indicate how frequently users navigated through those specific concepts and paths. For example, a thick, green line leading directly from the start to the correct destination suggests that most participants navigated the information architecture with ease, following a clear and intuitive path [238].

Pietrees offer a precise and insightful visualization of participants' navigation paths during a task, allowing to assess how participants build their individual conceptual models for the domain [238].

As suggested by Tullis and Albert [227], positive results in a UX study indicate success but it is by analyzing negative results – where users encounter difficulties – that we can uncover the weaknesses in our information architecture. Accordingly, this section focuses on discussing the negative results in each task, such as failure and wrong paths, since these highlight the most critical issues in the ontology. The pietree figures presented in this section are also available from Zenodo [14].

8.7.2 Task 1 - Target Concept: Whey

Fig. 8.13 showed that one participant incorrectly selected *cow milk* as the superclass for *whey*. However, *cow milk* is classified as a raw product, whereas whey is considered a by-product and for this reason could not be a subclass of *cow milk*. We believe other relationships between concepts in the ontology could provide clarity, such as `whey` `` `derives from' some `Cow milk'``. We believe that if participants had access to these additional relationships, they would have been less likely to misclassify whey under *cow milk*. Another common mistake was classifying whey under *Milk (processed) product*, likely influenced by AGROVOC's categorization. In AGROVOC, whey is incorrectly modeled as a processed product, whereas it should be classified as a by-product (the class milk by product in AGROVOC has no subclasses).

The rationale for categorizing whey as a by-product in APTO rather than as a processed product is based on its nature as a by-product of cheese, as noted by [239, 240]. In APTO, whey is modeled as a by-product with an added relationship, `whey `residue of' some `Cow cheese'`, to connect whey with its source product. We believe that in a real-world use scenario, the additional relationships provided by APTO would help prevent such misinterpretations of the concept.

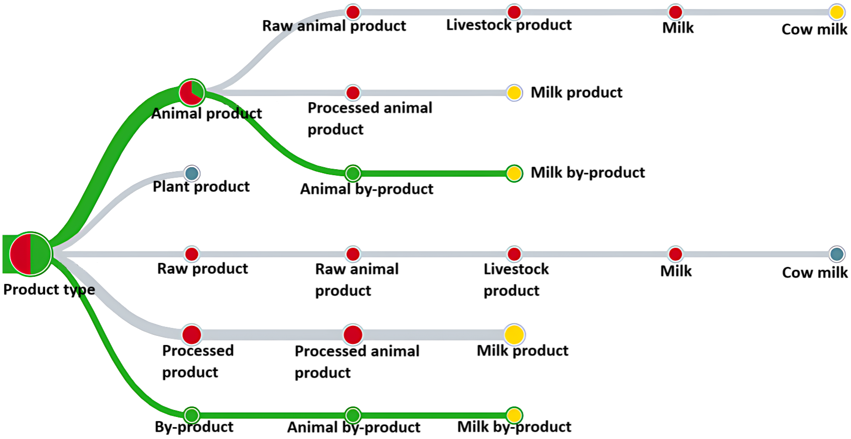


Figure 8.13 Pietree visualization for Task 1 - Classification of Whey. This task assessed whether participants could correctly classify whey as a milk by-product. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

8.7.3 Task 2 - Target Concepts: Soybean Meal and Wheat Bran

Fig. 8.14 reveals that participants frequently backtracked and explored multiple paths during the task. A notable observation is that many participants incorrectly classified *wheat bran* and *soybean meal* under *Cereal product*. This misunderstanding would be somewhat understandable if the task had only involved *wheat bran*, which is derived from wheat flour production. However, the task required participants to identify a common upper class for both *soybean meal* and *wheat bran*, and since soybean is not a cereal, *soybean meal* cannot be classified as a cereal product.

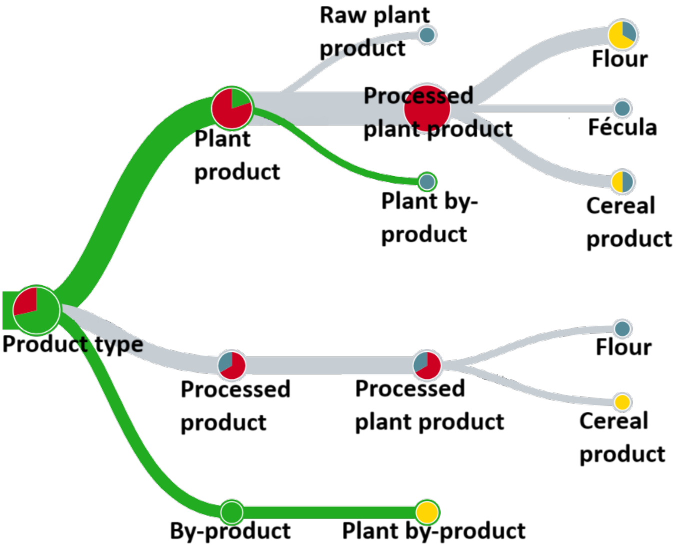


Figure 8.14 Pietree visualization for Task 2 – Classification of Soybean Meal and Wheat Bran. This task evaluated whether participants would classify these products as plant by-products. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

Most participants classified these products as processed products rather than by-products. To address this misconception, we have modeled relationships within the ontology to clarify the connection between processed products and the by-products they generate. Specifically, we used object properties to express that a by-product derives from another product, as follows:

```
Class: 'Soybean meal'
    SubClassOf: 'Plant by-product'
    residue_of: exactly 1 'Soybean oil'

Class: 'Wheat bran'
    SubClassOf: 'Plant by-product'
    residue_of: exactly 1 'Wheat flour'
```

In simpler terms, this expression defines that wheat bran is a plant by-product and explicitly models that it results from the processing of wheat flour. The object property residue_of indicates that the by-product is derived from another product.

8.7.4 Task 3 - Target Concepts: ‘Boi Magro’, ‘Boi Gordo’, ‘Vaca Gorda’, and ‘Vaca Leiteira’

Participants are rather precise in their navigation choices for this task since Fig. 8.15 shows that only two paths have been navigated through. Many participants who initially navigated down the wrong path backtracked upon reaching the class *Livestock product*. This suggests that they recognized the task’s requirement to locate the superclass representing the entire animal, rather than a product derived from or part of the animal, which are options under the *Livestock product* class. These participants then corrected their path and successfully identified the correct superclass (*Bovine*).

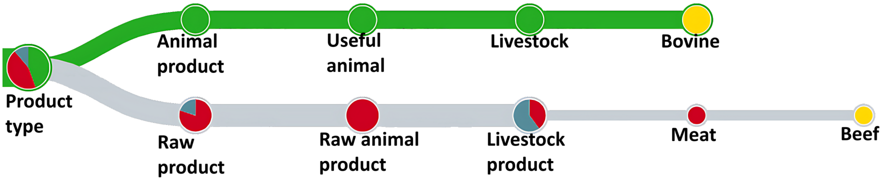


Figure 8.15 Pietree visualization for Task 3 – Classification of Boi Magro, Boi Gordo, Vaca Gorda, and Vaca Leiteira. This task assessed whether participants would correctly classify these terms under the superclass *Bovine*. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

Some participants classified the concepts under *Beef*, indicating they perceived a relationship between the concepts. Indeed, beef is a product derived from animals of the *Bovine* class. To capture this relationship within the ontology, we modeled it as:

```
Class: Beef
    SubClassOf: Meat
    derives_from: some Bovine
```

8.7.5 Task 4 - Target Concepts: Frozen Chicken and Chilled Chicken

In Task 4, some participants selected *Hive product* as the superclass, while others initially navigated to *Aquaculture product* but backtracked after realizing it was not the correct choice, as Fig. 8.16 shows. Neither of these classes is related to chicken, as they pertain to products derived from bees and aquatic animals, respectively. We believe this misunderstanding stemmed from the similarity of the upper class names in

Portuguese. The correct upper class, *'Produto da avicultura'*, closely resembles the names *'Produto da apicultura'* and *'Produto da aquicultura'*, which have similar syntactical structures and may have caused confusion.

Additionally, some participants navigated to the class *Processed animal product* but backtracked upon realizing it was not the correct upper class. Regarding the correct answers, all four valid paths were used by the participants, indicating that the polyhierarchical modeling of this class aligns well with the participants' conceptual understanding of the domain.

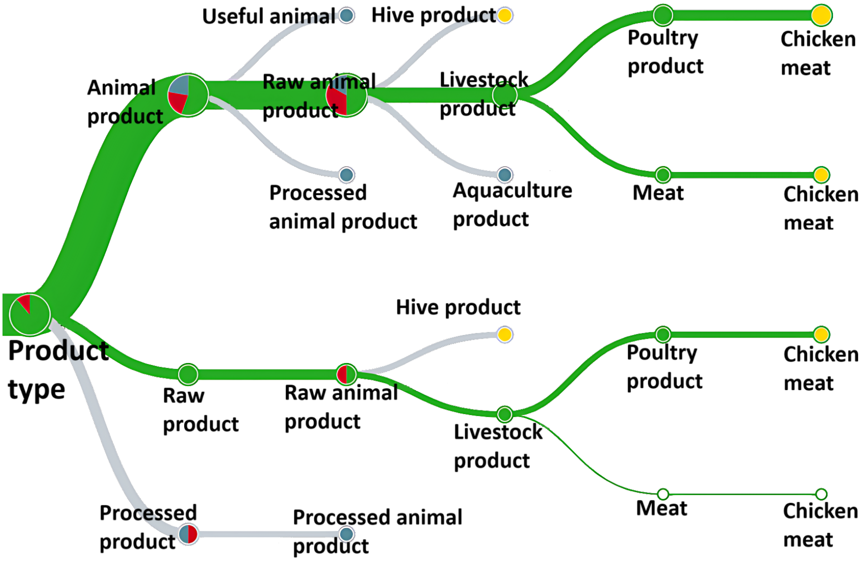


Figure 8.16 Pietree visualization for Task 4 – Classification of Frozen Chicken and Chilled Chicken. This task evaluated whether participants would correctly classify these items as raw animal products. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

8.7.6 Task 5 - Target Concepts: Dairy Drink, Yogurt, Powdered Milk, Pasteurized Milk, UHT Milk, and Cheese

Fig. 8.17 shows that in task 5 participants selected both of the correct paths defined for the task, indicating that the polyhierarchical modeling of this class aligns well with their conceptual understanding of the domain. However, one participant incorrectly selected *Cow milk* as the upper class. It is worth noting that hierarchical relationships in an ontology are often interpreted as 'is a' relationships. Thus, concepts like yogurt, dairy drink, and cheese would be incorrectly inferred as "Cheese is a Cow milk," in case the ontology was modeled this way.

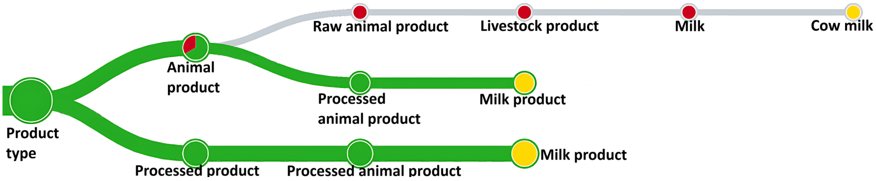


Figure 8.17 Pietree visualization for Task 5 – Classification of Dairy Drink, Yogurt, Powdered Milk, Pasteurized Milk, UHT Milk, and Cheese. This task assessed whether participants would classify these items under the superclass Milk Product. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

The task also included products such as Powdered Milk, Pasteurized Milk, and UHT Milk, which might raise questions about their classification. While these products are indeed types of milk, they undergo extensive processing that significantly alters their composition, classifying them as processed or even ultra-processed products. Since *Cow milk* has *Raw animal product* as its upper class in APTO, these processed milk products do not fit within this category, which represents unprocessed milk. We believe that the associative and hierarchical relationships defined in the full ontology could help prevent such misinterpretations.

In APTO, all the classes in Task 5 are defined as subclasses of *Milk product*. While it might seem natural to add a relationship indicating that a *Milk product* derives from *Cow milk*, this is not always the case. Some milk products are made from the milk of other animals, such as goats, as is the case with certain cheeses. Therefore, the relationship between a specific milk product and the type of milk it comes from is modeled at a more granular level, as shown below:

```
Class: 'Cow cheese'
    SubClassOf: 'Cheese'
    has_ingredient: some 'Cow milk'

Class: 'Goat cheese'
    SubClassOf: 'Cheese'
    has_ingredient: some 'Goat milk'
```

The relationship is set as 'some' rather than 'exactly 1' because all cheeses contain ingredients in addition to milk.

8.7.7 Task 6 - Target Concepts: Araucaria, Eucalyptus, and Pinus

Fig. 8.18 shows a modeling issue in APTO for task 6 regarding the classes *Araucaria*, *Eucalyptus*, and *Pinus*. In APTO, these classes were originally modeled as subclasses of either *Wood-producing plant* or *Cellulose-producing plant*, both of which are subclasses of *Useful plant*. However, most participants classified these classes as subclasses of *Forest product*.

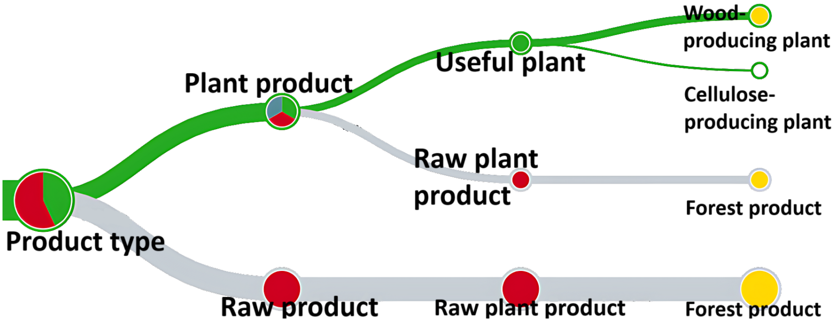


Figure 8.18 Pietree visualization for Task 6 – Classification of Araucaria, Eucalyptus, and Pinus. This task evaluated whether participants would classify these tree species under categories such as Wood-producing Plant or Forest Product. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

Initially, the *Forest product* class in APTO was designed to group raw products derived from trees, such as wood. We did not consider the trees themselves as forest products. However, after carefully analyzing the participants' choices, we recognized that these trees could indeed be considered forest products, given their role in forestry and their direct economic value as such. This led to revision of APTO to reflect a new hierarchy inferred from the participants' classifications, as shown below:

```
Class: 'Wood-producing plant'
    SubClassOf: 'Useful plant'
    SubClassOf: 'Forest product'


Class: 'Cellulose-producing plant'
    SubClassOf: 'Useful plant'
    SubClassOf: 'Forest product'
```

Of all the insights obtained from this study, we consider this one to be amongst the most valuable, as it allowed us to identify and correct a fundamental modeling issue in APTO.

8.7.8 Task 7 - Target Concepts: Andiroba almond, Cocoa bean, Cashew nut, and Coconut

Task 7 had no incorrect solutions, meaning all users successfully completed the task. However, the directness score of only 33%, as shown in Fig. 8.8, indicates that while users ultimately found the correct path, many initially navigated through incorrect paths before identifying the right one. Fig. 8.19 provides insight into these wrong paths.

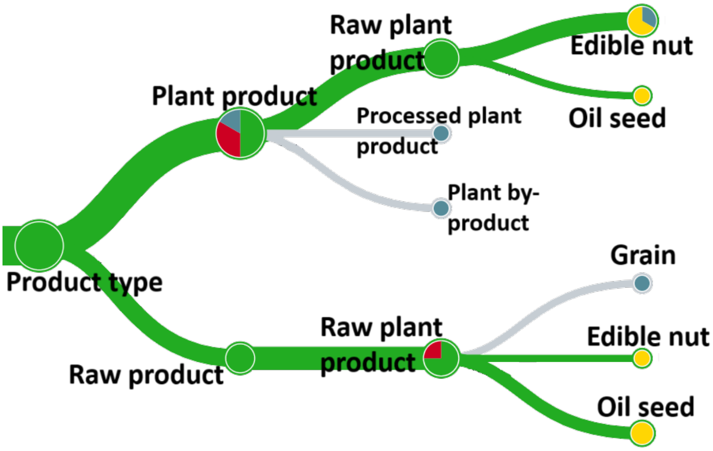


Figure 8.19 Pietree visualization for Task 7 – Classification of Andiroba almond, Cocoa bean, Cashew nut, and Coconut. This task assessed whether participants would classify these items as edible nuts or oil seeds. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

In the upper branch of the Pietree, we observe that some participants initially navigated to *Processed plant product* and *Plant by-product* before backtracking. This is encouraging, as it shows participants correctly recognized that these were not the appropriate superclasses, given that all target concepts in this task represent raw plant products.

In the lower branch of the Pietree, we see that some participants navigated to *Grain* and then backtracked. This reveals a higher-level ontological relationship between the target concepts, which are all types of nuts, and the class *Grain*, as both nuts and grains are types of seeds. Although this relationship is not currently modeled in APTO, it will be considered for inclusion in the next release.

8.7.9 Task 8 - Target Concepts: Pineapple pulp, Açaí pulp, Buriti pulp, Ceriguela pulp, and Cupuaçu pulp

Fig. 8.20 shows a common mistake in Task 8, which was the selection of the superclass *Fruit* as the correct answer. However, since *Fruit* is categorized under the superclass *Raw product*, this choice is inconsistent with the nature of *fruit pulp*. *Fruit pulp* is an extract derived from fruit, which means the ontological relationship between fruit and fruit pulp is one of part/whole, rather than a parent/child relationship, making it unsuitable to classify *fruit pulp* directly as subclass of *Fruit*.

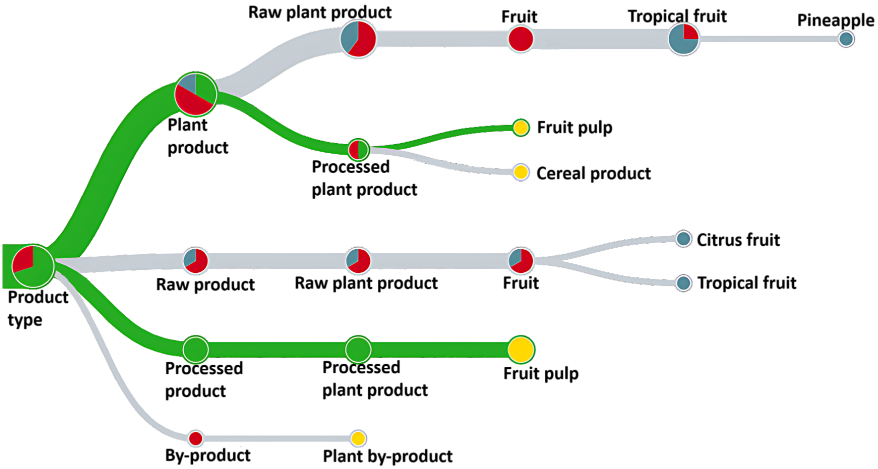


Figure 8.20 Pietree visualization for Task 8 – Classification of Pineapple pulp, Açaí pulp, Buriti pulp, Ceriguela pulp, and Cupuaçu pulp. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

Fruit pulps undergo various stages of processing to achieve the desired consistency, flavor, and other properties, which often results in a reduction of beneficial characteristics, such as dietary fiber content [241]. Despite the incorrect selection, it is clear that participants recognized the intrinsic relationship between fruit and fruit pulp. In APTO, we have used object properties to accurately model this relationship between fruit pulp and the fruit from which it is derived, as illustrated for the case of an specific fruit (Assai) below:

```
Class: 'Assai pulp'
    SubClassOf: 'Fruit pulp'
    derives_from: exactly 1 Assai
```

This type of relationship represents a recurring modeling problem in APTO and could benefit from the development of an Ontology Design Pattern (ODP) [242]. Specifically, the relationship between derived products (e.g., fruit pulp) and their source (e.g., fruit) requires a consistent and scalable approach that ensures semantic accuracy while minimizing manual effort. An ODP for this purpose could streamline the ontology development process through automation, as outlined below:

1. **Template Definition**: Create a reusable template or script that generates subclasses of Fruit pulp for each fruit in APTO.
2. **Automatic Restriction Assignment**: Automate the assignment of the derives_from exactly 1 [Fruit] restriction to each subclass, ensuring consistency in the representation of derivation relationships.
3. **Implementation with Rule-Based Tool**s: Employ a rule-based system or a scripting language (e.g., Python with OWL-RDF libraries or Protégé plugins) to apply the pattern systematically across all relevant fruit types in APTO.

This ODP will be adopted for future updates of APTO to support scalability in modeling this kind of relationship.

8.7.10 Task 9 - Target Concepts: Rice flakes and Corn flakes

In Task 9, participants mistakenly selected *Flour*, *Fiber*, and *Plant by-product* as the correct superclasses, as shown in Fig. 8.21. None of these classes have any relationship with *Corn flakes* and *Rice flakes*. However, none of these classes have a direct relationship with *Corn flakes* or *Rice flakes*. An interesting observation is that most participants followed the correct path up to *Processed plant product*, but then backtracked, unable to find a class they deemed appropriate for the task. To clarify the meaning of these classes and prevent such misunderstandings in the future, we have added definitions to these concepts in APTO.

8.7.11 Task 10 - Target Concepts: Cane sugar and Ethanol

Fig. 8.22 shows that *Plant by-product* was a popular choice among participants as the correct answer for this task. However, neither *Cane sugar* nor *Ethanol* are by-products (they are processed products).

Some participants also selected *Sugar cane* as the answer, indicating they recognize a relationship between these products and their source

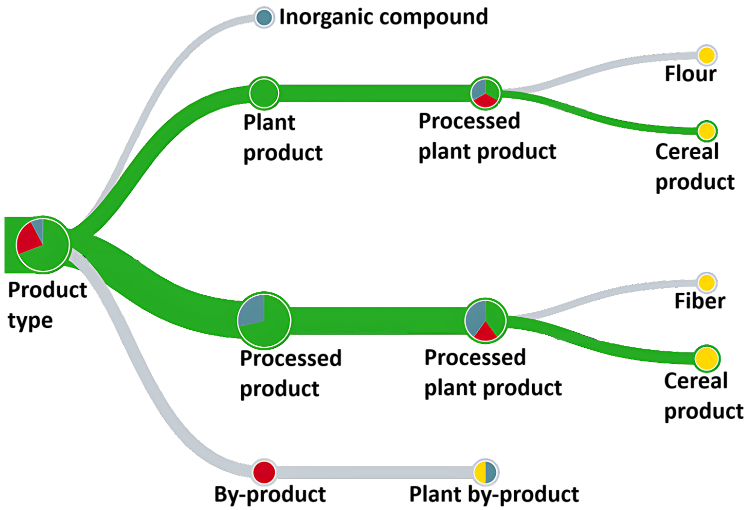


Figure 8.21 Pietree visualization for Task 9 – Classification of Rice flakes and Corn flakes. This task assessed whether participants would identify these items as processed cereal products. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

(sugar cane). However, this relationship cannot be classified as a parent/child relationship, since *ethanol* and *cane sugar* are not subtypes of sugar cane, but they are products derived from the processing of sugar cane. To accurately represent this relationship, we modeled it as follows:

```
Class: 'Cane sugar'
    SubClassOf: 'Sugar'
    derives_from: exactly 1 'Sugar cane'
```

```
Class: 'Ethanol'
    SubClassOf: 'Alcohol'
    derives_from: some ('Sugar cane' or 'Maize')
```

In the case of ethanol, it can also be derived from maize, which is why this was included in the `derives_from` relationship.

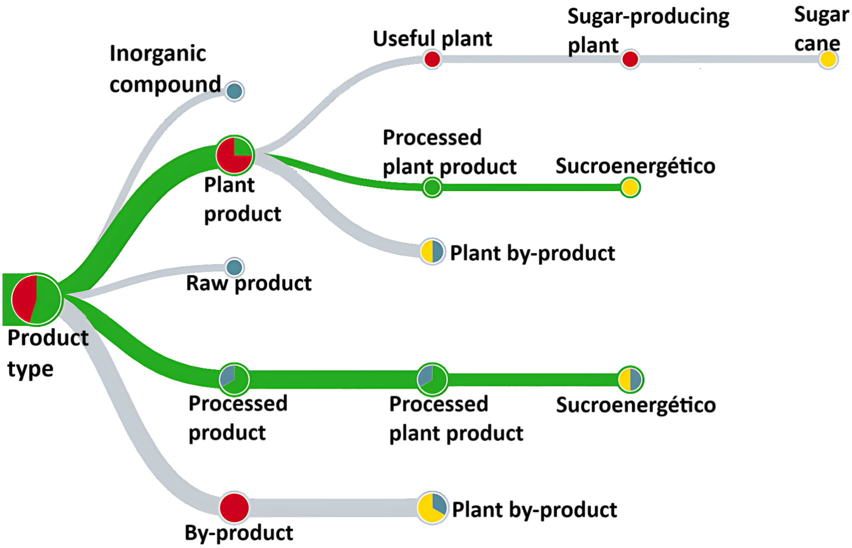


Figure 8.22 Pietree visualization for Task 10 – Classification of Cane sugar and Ethanol. This task evaluated whether participants would correctly associate these products with the newly introduced superclass Sucroenergético. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

8.7.12 Task 11 - Target Concepts: Babassu oil, Macauba oil, Buriti oil, Copaiba oil, Murumuru oil, and Pequi oil

The only incorrect answer for Task 11 was selecting the class *Oil seed*, as shown in Fig. 8.23. This choice suggests that the participant understood there is some relationship between the concepts. However, it is not a parent/child relationship, since oil undergoes several processing stages and cannot be considered a subtype of *Oil seed*, which is categorized as a raw product.

In answer to this, in APTO we have modeled the relationships between types of plant oil and the seeds they originate from, as illustrated in the example below:

```
Class: 'Soybean oil'
    SubClassOf: 'Plant oil'
    derives_from: exactly 1 Soybean

Class: 'Murumuru oil'
    SubClassOf: 'Plant oil'
```

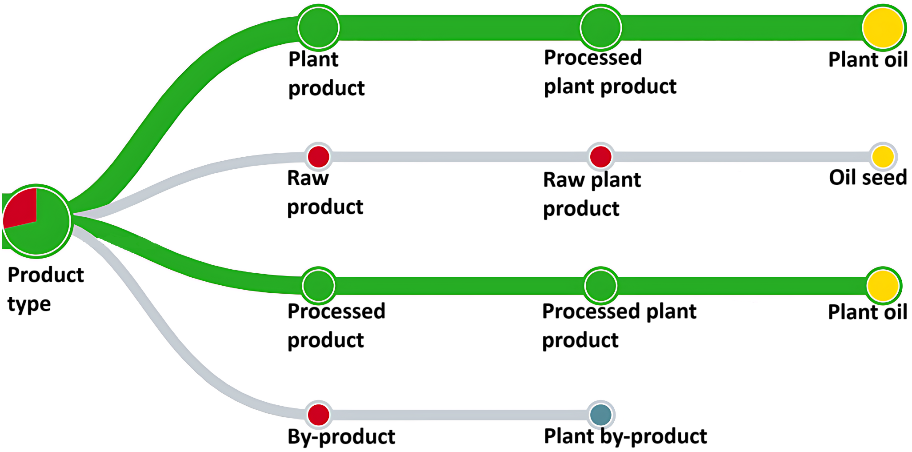


Figure 8.23 Pietree visualization for Task 11 – Classification of Babassu oil, Macauba oil, Buriti oil, Copaiba oil, Murumuru oil, and Pequi oil. This task assessed whether participants would classify these oils under Plant Oil. Node colors indicate user actions: green nodes represent navigation through correct paths; red nodes, incorrect paths; blue nodes, backtracking behavior; and the yellow node marks the selected destination. Line colors show navigation direction: green lines represent correct paths; gray lines represent incorrect ones. The size of the nodes and the thickness of the lines reflect the frequency with which participants navigated or selected them.

```
    derives_from: exactly 1 Murumuru
```

Upon analyzing this recurring relationship, we identified it as an ODP. Specifically, every subclass of `Plant oil` consistently includes a `derives_from` relationship to specify the source of the oil. For the next version of APTO, we will also automate this ODP.

8.8 Discussion

This study highlights the value of using the Tree Testing protocol to evaluate an ontology module, providing significant insights into participants' navigation paths and choices. These findings contribute to a deeper understanding of how users conceptualize the domain and interact with hierarchical structures in the ontology. However, several limitations emerged during the study, which provide opportunities for improvement and future exploration.

8.8.1 Advantages and Limitations of the Tree Testing Protocol for Ontology Evaluation

The Tree Testing protocol proved to be an effective tool for identifying usability issues in the ontology, particularly in evaluating hierarchical navigation. The Pietrees, introduced as part of the analysis, offered a detailed visualization of user navigation paths, which helped identify patterns of incorrect answers and backtracking behavior. For example, tasks with lower directness scores highlighted specific areas where participants struggled to find correct paths, revealing weaknesses in the ontology's structure.

However, a key limitation was the inability to fully understand participants' reasoning behind their navigation choices. While navigation paths provided quantitative data, they lacked context about why participants selected certain classes. Although open-ended questions were included in the questionnaire to address this limitation, participants provided minimal feedback. This suggests that complementary methods, such as interviews or think-aloud protocols, could provide richer insights into participants' thought processes and the conceptual challenges they face when navigating the ontology.

Another limitation was the exclusive use of hierarchical relationships in the prototype. Participants did not have access to associative relationships (expressed by object properties), which might have clarified the relationships between concepts. For instance, in Task 1, confusion about whether whey is a milk by-product or a processed product might have been mitigated if participants could see the additional relationship `whey `derives from' some `Cow milk'`. Including these relationships in future iterations of the study could improve participants' ability to make accurate classifications and align their understanding with the ontology's design. Thus, a different protocol should be used, since the tree testing protocol only works for hierarchical relationships.

8.8.2 Task-Specific Observations and Challenges

Some task designs may have influenced the results. For example, in Task 6, participants struggled to correctly classify *Araucaria*, *Eucalyptus*, and *Pinus*, possibly because the prototype did not reflect participants' mental models of these trees as forest products. Similarly, in Task 5, the classification of *Dairy Drink* was complicated by ongoing debates in Brazil regarding its quality and composition. These examples highlight the need for clearer definitions and targeted testing of new concepts, particularly those that are contentious or culturally specific.

8.8.3 Future Directions

Building on these findings, future studies should adopt a mixed-methods approach to ontology evaluation. Combining Tree Testing with qualitative methods — such as interviews or direct observation of users interacting with the ontology — could yield deeper insights into participants' decision-making processes and reasoning patterns.

In parallel, future research should also consider extending usability protocols beyond hierarchical structures to include associative relationships (i.e., object properties). While Tree Testing effectively captures users' understanding of class hierarchies, it does not support the evaluation of other semantic relations, such as `derives_from`, which are essential for interpreting provenance and conceptual derivation. Incorporating visual or interactive mechanisms that make these non-hierarchical relationships explicit may improve classification accuracy and provide a more comprehensive assessment of ontology usability.

Other protocols such as think-aloud and interviews could prove valuable in ontology usability studies, where errors may not arise solely from misunderstandings of hierarchical relationships. These protocols would allow users to verbalize the difficulties they encounter while interacting with the ontology and offer more detailed feedback on their navigation choices. However, these protocols tend to generate noisier data and are generally more complex to analyze than the method we used, as they involve interpreting qualitative responses.

Another avenue for future work is the systematic development and automation of ODPs to address recurring modeling challenges in APTO. For example, automating the assignment of `derives_from` relationships for derived products could streamline ontology development while ensuring semantic accuracy. Expanding the ontology's scope to include such patterns would also facilitate its integration with other vocabularies and enhance its usability for domain experts.

Finally, although this study focused on a specific ontology, APTO, which was designed for the Brazilian agricultural domain, the Tree Testing-based evaluation protocol has the potential to be applied in other domains. We find this protocol especially suited for ontologies that feature rich hierarchical structures. However, its effectiveness may vary depending on the depth, balance, and complexity of the class hierarchy of the ontologies being evaluated. Ontologies with relatively flat structures or heavily reliant on associative (non-hierarchical) relationships may require adaptations or complementary evaluation methods. Future research could explore the applicability of this approach across diverse ontology types and domains, helping establish broader methodological guidelines.

8.9 FAIR Assessment

As outlined in the ontology requirements elicitation in Section 6.4, we aimed to make APTO as FAIR as possible. To achieve this, we followed the ontology FAIR assessment methodology proposed by Amdouni et al. [151], which is integrated into AgroPortal. The O'FAIRe methodology provides an automated evaluation of an ontology's compliance with the FAIR principles through a structured set of 61 questions, primarily based on metadata descriptions. Since APTO is an agricultural ontology and is available in AgroPortal, implementing this method was a logical choice, as it is specifically tailored for ontologies hosted within AgroPortal.

At the time this assessment was conducted, AgroPortal's FAIR assessment tool was still in its Beta phase and displayed warnings indicating that it might not function properly. Additionally, out of the 61 questions in the FAIR assessment framework proposed by Amdouni et al. [151], 11 questions had yet to be implemented in the tool. These questions are represented by the gray bars in Fig. 8.24. During our evaluation, we encountered several issues when analyzing the assessment results. Rather than focusing on the positive scores, this section highlights these issues, as we believe they provide more valuable insights for the community. The APTO overall FAIRness score is shown in Fig. 8.24.

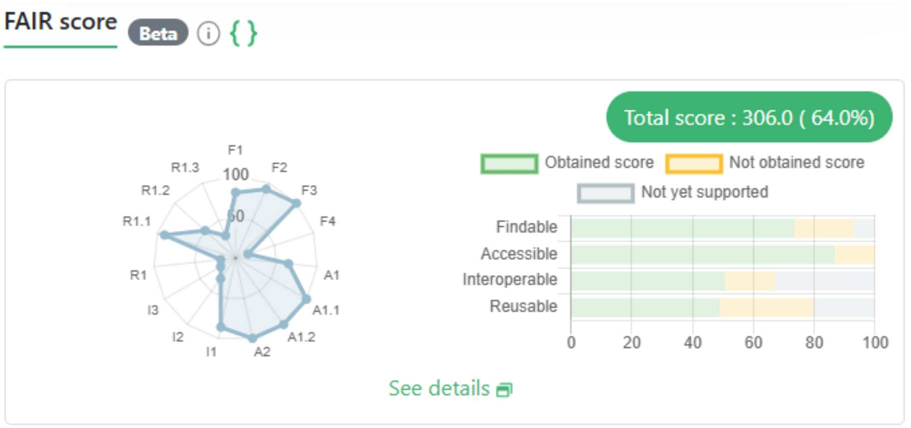


Figure 8.24 APTO FAIR Assessment on Agroportal.

Regarding principle F1, the ontology scored 33 points (80%), failing to meet only requirement F1Q2, which states:

> "Does the ontology provide an additional 'external' identifier, i.e., a guaranteed globally unique and persistent identifier assigned by an accredited body? If yes, is the external identifier a DOI?" [151].

Although our ontology has a DOI from Zenodo and a persistent identifier from W3ID, the question focused only on the DOI and classified the presented DOI as “External identifier is present but invalid,” reducing the score for this question to only three points. The reason for this classification remains unclear. Furthermore, the question itself presents certain issues, as it enforces the use of DOIs as unique identifiers, despite DOIs being a proprietary technology that does not guarantee perpetual persistence – DOIs can become broken over time. Given these limitations, we believe F1Q2 could benefit from further refinement to better reflect the realities of persistent identification in the context of ontologies.

For principle F2, the ontology scored 25 points (92%), failing to meet only one requirement: F2Q2, which recommends:

> “Is the ontology described with additional ‘MIRO should’ or ‘optional’ metadata properties?” [151].

This question suggests incorporating metadata properties such as omv:numberOfClasses, omv:numberOfProperties, omv:numberOfAxioms, and omv:numberOfIndividuals from the Ontology Metadata Vocabulary (OMV). However, the OMV namespace URI[3] is broken and does not resolve. While OMV is available on BioPortal, we opted not to reuse it to avoid integrating broken URIs into our ontology. This decision aligns with Linked Data principles, which emphasize the importance of persistent and resolvable identifiers.

For principle F3, APTO achieved a full score of 21 points (100%), whereas for F4, it scored only 4 points (16%). The lower score for F4 is due to APTO being indexed in only one ontology library, while question F4Q1 recommends indexing the ontology in at least three libraries. However, the definition of ”ontology libraries” in this context is unclear. The assessment suggests using the schema:includedInDataCatalog property to indicate the libraries, which could imply ontology catalogs. However, question F4Q2 explicitly refers to ontology repositories such as AgroPortal, yet it also uses the schema:includedInDataCatalog property while requiring the ontology to be available in at least five repositories. This ambiguity raises questions about the distinction between ontology catalogs and repositories in the evaluation criteria.

For principle A1, APTO scored 28 points (65%). The first question, A1Q1, evaluates whether the ontology provides resolvable URIs. Although the tool correctly recognized both the namespace URI and the Ontology Zenodo DOI as resolvable, it assigned a score of 0, which we believe is an error.

Regarding A1Q3, which assesses whether the ontology URI supports content negotiation, APTO scored 15 points. This is because it supports only five negotiation formats, whereas at least eight are recommended for full compliance. This will be analyzed as future improvements.

---

[3] `http://omv.ontoware.org/2005/05/ontology`

For principle A1.1, APTO achieved a perfect score of 28 points (100%), as well as 22 points (100%) for A1.2 and 20 points (100%) for A2. These strong results highlight the ontology high compliance with accessibility.

For principle I1, APTO scored 38 points (86%). not reaching 100% was due to the absence of additional format serializations in the ontology metadata. However, after updating the metadata to include this information, the FAIR Score on AgroPortal remained unchanged. This suggests that the score does not update when the ontology is modified in the portal.

For principle I2, APTO scored only 10 points (31%), primarily due to question I2Q2, which assesses the reuse of other FAIR vocabularies. Although APTO integrates many FAIR vocabularies, this was not specified in the ontology metadata using dct:relation, as recommended by the assessment tool. Even after updating the metadata, the score did not change, mirroring the issue observed in principle I1.

Similarly, question I2Q6 evaluates relationships with other FAIR vocabularies. Despite adding relevant metadata, the score remained unchanged. For question I2Q7, which concerns the use of FAIR metadata standards to describe ontology metadata, we incorporated elements from FAIR metadata standards such as Dublin Core and Schema.org. However, as with the previous questions, the score did not update. For principle I3, APTO achieved a full score of 7 out of 7 points, as implemented in the FAIR assessment tool.

For principle R1, APTO scored 6 out of the 8 possible points for the implemented questions. The deduction was due to question R1Q2, which assesses whether the ontology provides information about its hierarchy. The assessment tool recommends using mod:obsoleteParent and mod:maxDepth, but we see no justification for including these properties.

The tool does recognize owl:SubClassOf as a valid property for describing hierarchy. However, since APTO only includes this single property rather than the three recommended, the assessment concluded that the ontology lacks sufficient hierarchy representation. We believe that enforcing the use of these additional properties is unnecessary.

For principle R1.1, APTO scored 34 points (91%) due to the absence of the cc:morePermissions property. Instead, we used dct:accessRights and dc:license to provide information about permissions. We believe these properties sufficiently cover all necessary permission details, making cc:morePermissions unnecessary.

For principle R1.2, the ontology scored only 19 points (50%) due to the absence of information about its developers (question R1.2Q1) and provenance details (question R1.2Q2). Although we updated the ontology to include this information, the score was not updated.

Finally, for principle R1.3, the ontology scored only 11 points (30%) due to the absence of endorsements from organizations (R1.3Q1) and its lack of inclusion in a specific group within AgroPortal. The latter issue was addressed by adding the ontology to the AgroTermos group. However, as with previous cases, the tool did not update the score.

This assessment provided valuable insights into the strengths and weaknesses of APTO's FAIR compliance, allowing us to identify and address missing critical metadata, such as provenance information and relationships with other FAIR vocabularies. By updating the ontology metadata accordingly, we improved its accessibility, interoperability, and reusability. However, a significant drawback of the evaluation process was the apparent failure of the FAIR assessment tool to update scores after metadata modifications, preventing an accurate reflection of the improvements made. Additionally, some scoring criteria, such as the enforcement of certain properties (e.g., mod:obsoleteParent, mod:maxDepth, and cc:morePermissions), seemed unnecessarily rigid, limiting flexibility in ontology design. Despite these issues, the assessment helped enhance the ontology's quality and alignment with FAIR principles, reinforcing its value for the research community.

8.10 Ontology Publication

In alignment with semantic web principles, APTO has been implemented in OWL and is publicly available on AgroPortal, which also provides a dedicated SPARQL endpoint for querying[4]. The ontology's persistent URI[5] is configured to support content negotiation by redirecting to AgroPortal. As a result, human users accessing the URI through a browser are presented with a user-friendly HTML interface, while machines receive data in machine-readable formats such as RDF/XML or JSON-LD. Furthermore, all released versions of the ontology are archived both on AgroPortal and on Zenodo[6], ensuring long-term accessibility and version tracking.

8.11 Final Considerations

We believe that combining the usability evaluation with the FAIR assessment significantly contributed to enhancing the accuracy and overall quality of APTO. The usability evaluation, through the three testing protocols, provided valuable feedback on how effectively users could interact with and understand the ontology, ensuring that it met practical needs and expectations. Meanwhile, the FAIR assessment allowed us to systematically refine the ontology's metadata, making it more accessible, interoperable, and reusable.

---

[4] `https://agroportal.lirmm.fr/ontologies/APTO?p=sparql`

[5] `https://w3id.org/APTO#`

[6] `https://doi.org/10.5281/zenodo.13271285`

CHAPTER 9

Knowledge Graph for Agricultural Price Data

Chapter Overview

This chapter presents the C4AI-KGAP, which is the result of the final stage of the Data Semantic Annotation process (Section 2.5). It details the transformation of datasets from CEPEA, CONAB, and Ipea into RDF, enabling cross-source data integration. The KG Structure (Section 9.3) follows the Almes Core model [131]. The Metadata KG describes datasets using dcat:Dataset, linking product types to APTO and locations to GeoNames. The Data KG represents price observations as blank nodes, denoted with alm:hasObservation. Custom Python scripts using rdfLib were used to generate RDF triples, overcoming OpenRefine's limitations in handling blank nodes. GeoNames data was extracted via API, and price data was mapped to RDF triples.

9.1 Knowledge Graph Design Methodology

The design and implementation of the C4AI-KGAP were guided by key principles and practices established in the KG engineering literature. While this chapter focuses on the implementation details of the KG, it is important to reflect on the underlying methodological assumptions and choices that informed its construction.

9.1.1 General Principles

The KG was developed following the core principles of Linked Data and Semantic Web technologies, as originally articulated by Berners-Lee [243] and formalized through W3C standards such as RDF, RDFS, OWL, and SPARQL [39, 118, 119]. These principles emphasize:

- the use of dereferenceable URIs to uniquely identify resources;

- the modeling of knowledge as RDF triples using well-defined vocabularies;
- interlinking of data from heterogeneous sources;
- enabling both human and machine interpretation of data semantics.

9.1.2 Methodological Approach

Methodologically, the KG design followed a modular and layered architecture, separating metadata (descriptive information about datasets) from observation data (price values over time), and aligning these layers to domain-specific and external ontologies (e.g., APTO, GeoNames).

Our approach fits within the design science research paradigm discussed in Chapter 1, while also drawing from structured methodologies in Knowledge Graph Engineering (KGE). KGE is an emerging discipline that proposes systematic processes for KG construction, typically involving stages such as requirement analysis, modeling, implementation, and maintenance [244, 245].

This chapter presents C4AI-KGAP, positioned within the theoretical framework of the Knowledge Graph Lifecycle proposed by [246]. This model delineates the processes required for building and maintaining high-quality knowledge graphs and includes stages such as creation, hosting, curation, and deployment.

Knowledge Creation

The Knowledge Creation stage involves generating RDF triples from raw data sources and linking them with ontologies or external vocabularies. Creation can be manual or (semi-)automated, often using mappings or custom transformation logic [246]. In this case study, this included:

- Mapping metadata to Almes Core using Python and RDFLib.
- Metadata records were instantiated as dcat:Dataset entities and populated using Almes Core.
- Associating products with APTO ontology classes and locations with GeoNames URIs.
- Representing price observations as blank nodes linked to dataset metadata via the alm:hasObservation property.

Our choice to implement custom scripts instead of using declarative mapping languages like RDF Mapping Language (RML) as implemented by the authors [246] was motivated by the need for fine-grained control over blank nodes, nested structures, and hash-based URI generation.

Knowledge Hosting

Knowledge Hosting refers to the storage, publication, and long-term accessibility of a KG, typically through repositories, Linked Data platforms, or triple stores that support querying, versioning, and interoperability [246]. In the case of C4AI-KGAP, the KG was hosted on the Platform Linked Data Netherlands (PLDN), which provides infrastructure for storing RDF data and executing SPARQL queries over triple stores. Additionally, all RDF components of the KG – including metadata, price observations, agricultural product ontologies, and geographic references – were serialized in Turtle format and published in open-access repositories such as Zenodo and GitHub. This strategy promotes persistent accessibility, facilitates citation and reuse, and aligns with FAIR principles by ensuring the findability and long-term availability of the KG artifacts.

Knowledge Curation

Knowledge Curation encompasses activities such as data cleaning, enrichment, assessment, and validation to ensure the accuracy, completeness, and semantic richness of the KG [246].

Application to C4AI-KGAP:

- Cleaning: We cleaned input data to fix malformed entries (e.g., date formats like yyyy.mm) and resolved identifier mismatches (e.g., merged cells in CONAB data) in a previous stage of this case study (refer to Section 4.4).
- Assessment: Internal tests ensured that the metadata conformed to Almes Core and that price entries were consistently typed and linked.
- Enrichment:
  - GeoNames enrichment was implemented by calling the GeoNames API for each location and appending gn:name and gn:geonamesID triples.
  - Semantic enrichment of products was achieved by mapping terms to APTO classes using string matching and manual curation.
- While SHACL or OWL reasoning was not yet implemented for validation, this work lays the foundation for future integration of constraint-checking tools like VeriGraph or RDFUnit.

Knowledge Deployment

Knowledge Deployment involves making the KG usable for end applications, such as search, question answering, analytics, or decision support systems [246].

C4AI-KGAP is queryable via RDF using tools like SPARQL for answering analytical questions such as:

- "What was the average price of beef in São Paulo between 2010 and 2020?"
- "Which products had the highest price volatility in a given period?"

This capability directly supports use cases seen in agricultural economics and policy, including cross-source price comparisons, market analysis, and temporal trend assessments.

By aligning the construction of C4AI-KGAP with the stages of the lifecycle model from [246], this chapter situates the work not only as a practical artifact but also as a theoretically grounded contribution to semantic knowledge engineering. The following sections detail the pipelines, tools, and methodologies employed at each stage of the lifecycle, illustrating how C4AI-KGAP adheres to established best practices in Knowledge Graph construction and management.

9.2 Use and Applicability of Price Index Data

The most common uses of price index data include time series analysis, as demonstrated in [247], and price comparisons across different regions in Brazil [248]. Price data also plays a crucial role in communication media broadcasts, as it is disseminated daily to inform the public about commodity prices. This information is vital for a wide array of stakeholders, including traders, supermarket chains, farmers, and others involved in the agricultural market chain. In Brazil, many media outlets, such as Globo Rural [249], Reuters [250], and others, publish this data on a daily basis, reflecting its importance to the public and industry players. Based on these common use cases, C4AI-KGAP was designed to support queries that align with the typical analytical tasks performed by users, as observed in the referenced studies. These include real-world needs such as exploring time series trends and comparing agricultural prices across different regions.

9.3 Knowledge Graph Structure

The conceptual model presented in Fig. 9.1 represents the structure of C4AI-KGAP. The entire KG structure is based on the Almes Core data model [131].

Starting with the `Metadata KG`, it includes a description for each dataset, or more specifically, their `metadata records`. Each record contains the attributes displayed in Fig. 9.1. Each dataset is identified as a `dcat:Dataset`, and unique URIs for each dataset were generated using hashes derived from files stored on GitHub, as in the example below:

```
<https://github.com/Filipi-Soares/MetaID/blob/main/ID.ttl#L7> a
    ↪ dcat:Dataset ;
```

Finally, we generated the Metadata Graph in Turtle, and made it available on Zenodo [15]. The metadata fields `alm:productType` and `alm:productGroup` in the Metadata KG are instantiated with classes from APTO, implying that the values for these fields are URIs from APTO. A corresponding KG in Turtle format has also been generated for APTO and is available on the same Zenodo repository [15]. Similarly, `sdo:location` in the Metadata KG is instantiated with GeoNames classes. A subset of the GeoNames ontology was extracted to represent all Brazilian regions (i.e., cities and states) included in the KG. The data was converted to Turtle format and is available on Zenodo [15].

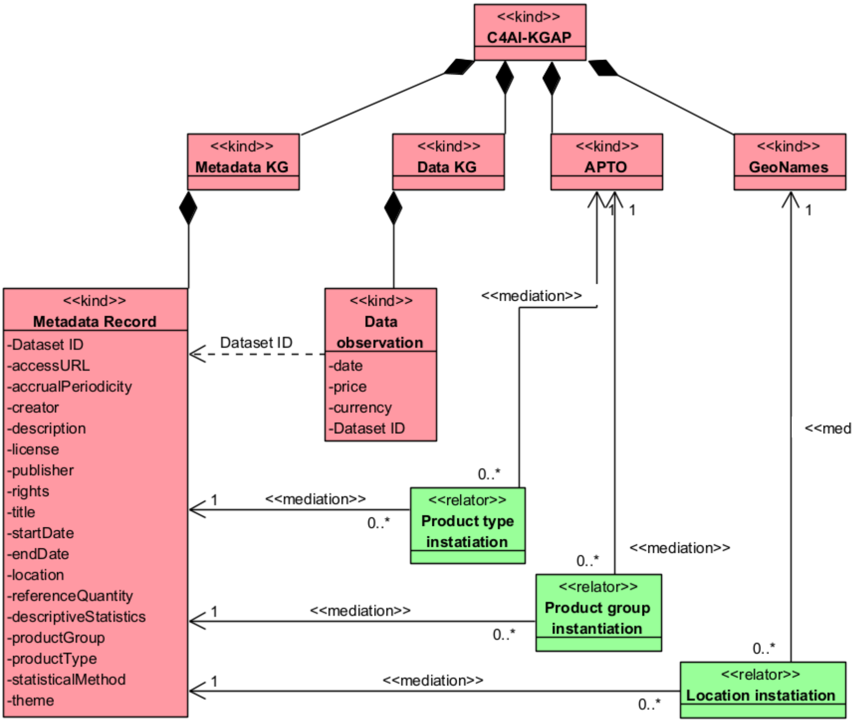


Figure 9.1 C4AI-KGAP Conceptual Model.

For the Data KG, each data point included attributes such as date, price, and currency, following the Almes Core recommendations [131]. The data points were modeled as RDF blank nodes of a Dataset ID from the Metadata KG, as illustrated in the example below:

```
<https://github.com/Filipi-Soares/MetaID/blob/main/ID.ttl#L102> alm:
    ↪ hasObservation [ sdo:currency "BRL"^^xsd:string ;
        sdo:date "2006-03-16"^^xsd:date ;
        sdo:price "24.86"^^xsd:float ],
```

In cases data was published as an interval, such as CONAB's Weekly Prices, the structure was adapted to use `sdo:startDate` and `sdo:endDate` instead of `sdo:date`. The Data KG has also been published on Zenodo [15].

Regarding the conversion of tabular data into RDF, we initially experimented with OpenRefine[1] and its RDF extension. However, we found it has limitations for handling data structures that require blank nodes.To address this issue, we developed custom Python notebooks using rdfLib to perform these conversions more effectively.

9.4 Creating the Metadata Graph

We designed a Python script (named metadata_converter.py in the Zenodo repository [17]) to convert the metadata CSV file previously generated and described in Section 4.4.1 (available on Zenodo [7]) into an RDF graph, which was then serialized into Turtle format. The script as shown in Fig. 9.2 works as follows.

The script begins by reading a CSV file containing metadata using the `pandas` library (`df = pd.read_csv`). Each row in this dataset represents a resource, such as a dataset with its associated properties.

Next, it initializes an RDF graph (`g = Graph()`) using `rdflib`, a Python library for working with RDF data. This graph stores the triples (subject-predicate-object relationships) for each resource.

To ensure semantic consistency, various namespaces are defined and bound to the graph (ALM, DCAT, SDO, XSD, DCT).

The script then iterates through each row in the CSV dataset as follows:

- A Uniform Resource Identifier Reference `URIRef` is generated for the resource, which acts as the subject of the RDF triples.
- Several RDF triples are added to the graph, mapping the resource's properties to RDF predicates, such as dct:title, dct:description, sdo:startDate, etc. Literal values such as dates and strings are converted using the appropriate datatype (e.g., XSD.date for dates and XSD.string for strings).

Once all the triples are added to the graph, it is serialized in the Turtle format.

Subsequently, we published this metadata file on GitHub [16] and generated actual URIs for each resource to replace the "https://example.org/resource/" URIs. These URIs have been generated using fragment identifiers (commonly referred to as hashes) to refer to specific parts of the Turtle file. In this case, the GitHub-hosted file uses fragment

---

[1] `https://openrefine.org/`

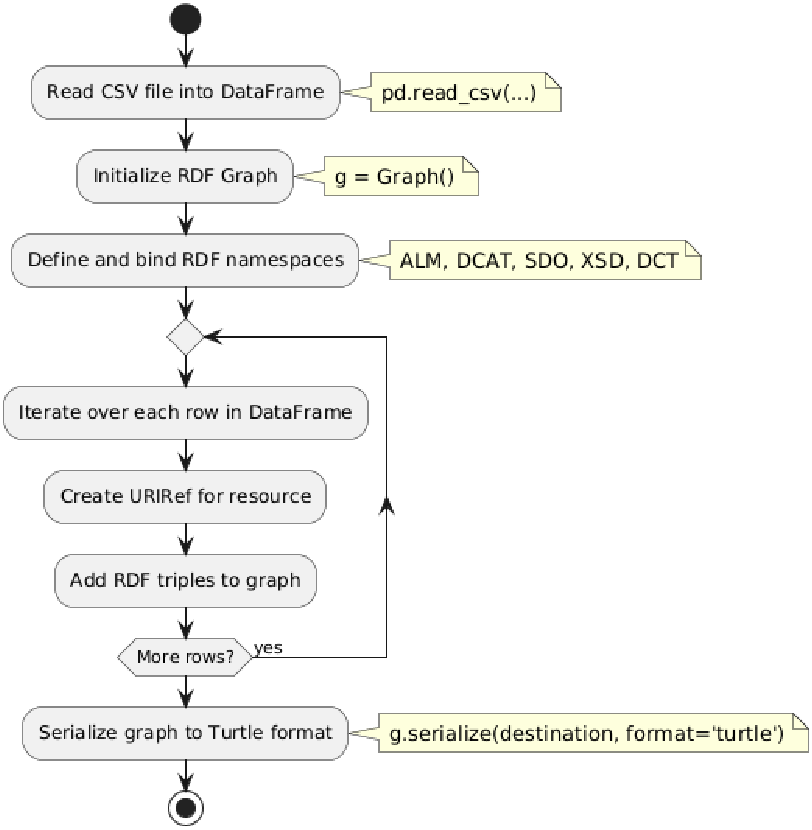

Read CSV file into DataFrame
pd.read_csv(...)
Initialize RDF Graph
g = Graph()
Define and bind RDF namespaces
ALM, DCAT, SDO, XSD, DCT
Iterate over each row in DataFrame
Create URIRef for resource
Add RDF triples to graph
More rows?
yes
Serialize graph to Turtle format
g.serialize(destination, format='turtle')

Figure 9.2 Metadata Converter Pipeline.

identifiers like #L7 to indicate a particular line within the file, in this case line 7. This makes it possible to point to specific entities within the document.

9.5 Extracting GeoNames Data for Locations

As mentioned earlier, the `sdo:location` field in the Metadata KG was populated with GeoNames URIs. To ensure that users querying the KG can view the location names associated with each URI, we extracted data from the GeoNames ontology via their API. This extraction was limited to the specific locations included in the KG, rather than the entire GeoNames ontology, and did not include the location hierarchy.

To achieve this, we developed a Python script (`GeoNames.py` Zenodo [17]), which functions as shown in Fig. 9.3. This script extracts GeoNames information for the geographic locations provided in the dataset, enriches the data with GeoNames metadata, and converts it into RDF/Turtle format for integration into the KG.

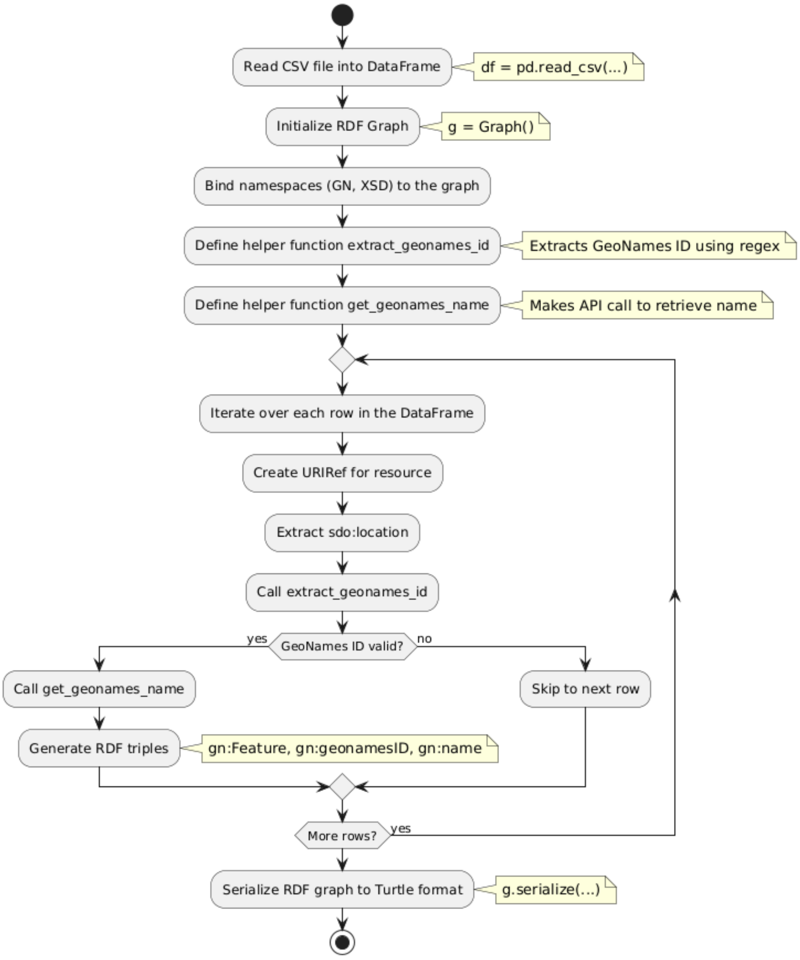

Read CSV file into DataFrame
df = pd.read_csv(...)
Initialize RDF Graph
g = Graph()
Bind namespaces (GN, XSD) to the graph
Define helper function extract_geonames_id
Extracts GeoNames ID using regex
Define helper function get_geonames_name
Makes API call to retrieve name
Iterate over each row in the DataFrame
Create URIRef for resource
Extract sdo:location
Call extract_geonames_id
yes
GeoNames ID valid?
no
Call get_geonames_name
Skip to next row
Generate RDF triples
gn:Feature, gn:geonamesID, gn:name
More rows?
yes
Serialize RDF graph to Turtle format
g.serialize(...)

Figure 9.3 Geonames Converter Pipeline.

The script starts by reading a CSV file into a pandas DataFrame, which contains the data from the Metadata KG, including geographic location information.

Next, an RDF graph is initialized to store RDF triples that will be created during the execution of the script.

To support semantic representation, the namespaces GN (representing GeoNames ontology) and XSD (for datatypes) are defined and binded to the graph using `g.bind`.

The script defines a helper function, `extract_geonames_id`, which uses regular expressions to extract the GeoNames ID from the location field in each row. This function identifies numerical IDs embedded within URL patterns that point to GeoNames resources.

Another helper function, `get_geonames_name`, takes the extracted GeoNames ID and sends an API request to the GeoNames service using `requests.get`. If the request is successful, it parses the JSON response to extract the geographic feature's name.

The script then iterates over each row in the dataset. For every row, a `URIRef` is created to represent the described resource. It extracts the sdo:location field and calls `extract_geonames_id` to retrieve the GeoNames ID.

If a valid ID is found, the script uses `get_geonames_name` to fetch the geographic name and then constructs RDF triples to represent the geographic feature.

For each valid GeoNames ID, the script generates the following RDF triples:

- The geographic feature is typed as a `gn:Feature`.
- The GeoNames ID is linked to the resource using `gn:geonamesID`.
- The name of the geographic feature is linked using `gn:name`.

Finally, the RDF graph is serialized to a Turtle file, storing the enriched geographic metadata for further use or integration.

9.6 Creating the Data Graph

The dataset, initially provided in CSV format (see Section 4.4.1, Zenodo [7]), contained four fields: `metadata_id`, date, price, and currency. To transform this data into RDF, we developed a Python-based pipeline, summarized in Figure 9.4.

The pipeline is divided into two main steps. First, a script replaces the ID values[2] in the `metadata_id` column with the corresponding URIs previously generated for the resources in the Metadata Knowledge Graph. The script loads the main dataset and the mapping file, builds a dictionary of ID-to-URI correspondences, replaces the original identifiers in the dataset, and saves the updated file for the next stage.

The second step focuses on transforming each row of the updated dataset into RDF. A new script loads the enriched CSV and initializes an RDF graph using the `rdflib` library. Namespaces from ALM, Schema.org (SDO), and DCAT are defined and bound to the graph to ensure consistent use of vocabulary.

Each row in the dataset is processed as an individual observation. For every observation, a blank node is created and annotated with RDF

---

[2] The metadata_id field was previously populated with strings for the IDs such as `IpeaFumo`, `CepeaBoiGordo`, `ConabAcucarAM`, etc.

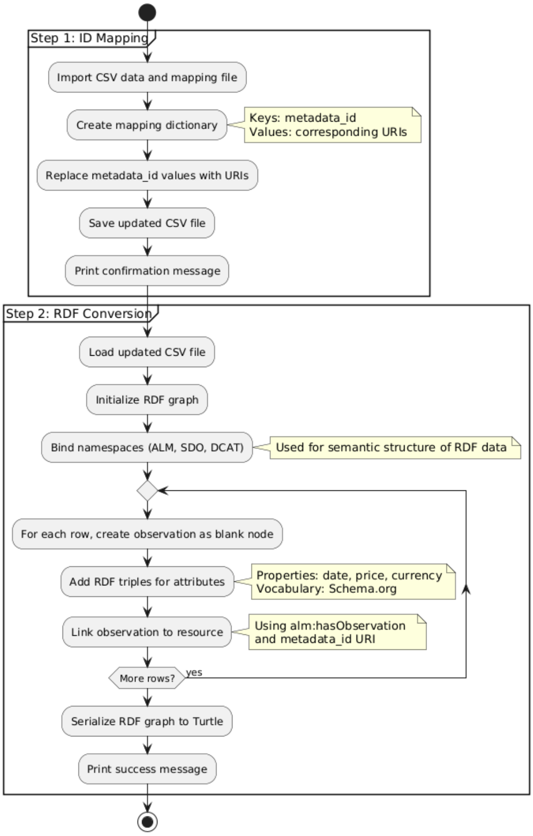

Step 1: ID Mapping
Import CSV data and mapping file
Create mapping dictionary
Keys: metadata_id
Values: corresponding URIs
Replace metadata_id values with URIs
Save updated CSV file
Print confirmation message
Step 2: RDF Conversion
Load updated CSV file
Initialize RDF graph
Bind namespaces (ALM, SDO, DCAT)
Used for semantic structure of RDF data
For each row, create observation as blank node
Add RDF triples for attributes
Properties: date, price, currency
Vocabulary: Schema.org
Link observation to resource
Using alm:hasObservation
and metadata_id URI
More rows?
yes
Serialize RDF graph to Turtle
Print success message

Figure 9.4 Data Converter Pipeline.

triples that describe its attributes: `date`, `price`, and `currency`. These values are linked using Schema.org predicates and typed according to the appropriate XML Schema datatypes (e.g., `xsd:date`, `xsd:float`, `xsd:string`).

To complete the model, each observation is linked back to its related metadata resource using the `alm:hasObservation` property and the URI from the `metadata_id` field. The final RDF graph is then serialized into Turtle format and saved to a file, ready for integration into the broader Knowledge Graph.

Both scripts are implemented in a single Python notebook named `data_converter.ipynb`, available on Zenodo [17].

9.7 Future Work

A challenge not currently addressed in our KG is how to manage price revisions. Agricultural price indexes are often revised to reflect market conditions more accurately, such as adjustments for inflation or deflation. For instance, historical prices may be updated to account for changes in purchasing power over time [251]. Future work should focus on developing a distributed application to monitor these revisions and create new entries in the KG. To represent these revisions correctly, each should be assigned a distinct URI, as the current use of blank nodes does not allow for differentiation between the original data and its revisions.

For future developments, the entire semantic interoperability framework could be automated using an integrated pipeline. This would require all participating institutions to agree on providing data formatted according to Almes Core via an API. Using a workflow tool like Apache Airflow, the pipeline could potentially automate periodic tasks such as data extraction, transformation, RDF conversion, and importation to the KG. Each step would run automatically when new datasets become available, ensuring that the KG remains continuously updated and accurate.

CHAPTER 10

Querying and Validating the Knowledge Graph Using LLM

Chapter Overview

This chapter marks the final stage of the application of the semantic interoperability framework described in Section 2.6, where we validate the functionality of the KG through SPARQL queries. By integrating data from three distinct sources, the KG enables cross-source analysis and comparison – capabilities that were previously not possible. Given the KG's complexity, we leveraged generative AI to automate SPARQL query generation from natural language inputs. In Section 10.2, we present a series of AI-generated SPARQL queries and analyze their results. Finally, in Section 10.5, we evaluate the KG's FAIRness, comparing its assessment with that of the original datasets to measure the impact of the framework on fostering data findability, accessibility, interoperability, and reusability.

10.1 Implementation of a Triple Store

To integrate the graphs into a single triple store, we merged the graphs within a single environment using PLDN. PLDN is a collaborative initiative that promotes the use and application of Linked Data technology in the Netherlands, and it provides a platform for publishing, sharing, and querying linked data, enabling the creation of a more interconnected data ecosystem. The platform allows organizations to make their data accessible and interoperable by converting it into machine-readable formats such as RDF and making it available through SPARQL endpoints [252, 253].

The PLDN platform supports the seamless integration of different datasets into a single triple store, which is particularly useful for projects like this that involve multiple data sources. By using PLDN, we were able to store, manage, and query the integrated knowledge graphs in a unified environment, ensuring that the data from various sources (e.g., GeoNames, APTO, and the datasets used in this study) could be queried together efficiently.

Each graph in this study is implemented as a **named graph** in PLDN:

- **graph:data**: The graph containing data from CEPEA, Ipea, and CONAB: `https://data.pldn.nl/FilipiSoares/AgriPrices/graphs/data`.
- **graph:metadata**: The graph containing the metadata descriptions of the datasets: `https://data.pldn.nl/FilipiSoares/AgriPrices/graphs/metadata`.
- **graph:apto**: The graph containing APTO: `https://data.pldn.nl/FilipiSoares/AgriPrices/graphs/apto`.
- **graph:geonames**: The graph containing extracted GeoNames data: `https://data.pldn.nl/FilipiSoares/AgriPrices/graphs/geonames`.
- The parent graph C4AI-KGAP, which encompasses all of the above: `https://data.pldn.nl/FilipiSoares/AgriPrices`.

In total, the integrated Knowledge Graph contains 162,305 statements, representing the diverse datasets used in this study.

10.2 Generating Queries with an LLM

As highlighted in [254], generative AI can support users unfamiliar with SPARQL on generating queries to retrieve data from KGs. The use of generative AI to create SPARQL queries from natural language inputs is not new, with numerous approaches tested in recent years. Avila et al. [255] conducted experiments testing ChatGPT's ability to answer natural language questions over KGs. Rony et al. [256] proposed SGPT, which uses Transformer encoders to map syntactic features of user questions to their corresponding SPARQL query representations. Similarly, Kovriguina [257] developed SPARQLGEN, a method for generating SPARQL queries using LLMs. Their approach involves augmenting LLMs with a KG fragment and question-subgraph query examples to generate the appropriate queries via structured prompts.

In this dissertation, our approach draws some ideas from these studies. However, the aim of using ChatGPT in this case study is not to evaluate its query-generating capabilities but to apply it in a real use case. Thus, our focus is on analyzing the structure of the generated queries and the results they produce, rather than the capabilities of LLMs in generating SPARQL queries, as discussed by [254, 255, 256]. Our method aligns closely with the work of Kovriguina [257], where

ChatGPT was trained by providing prompt-based interactions to supply information about the KG's architecture.

In our study, we used Semantic Web Senior, a specialized GPT developed from ChatGPT 4.0, the latest version available. This GPT is specifically trained with knowledge in RDF, RDFS, OWL, SPARQL, Linked Data, SHACL, Web Annotations, and Ontology Engineering, enhancing its ability to perform complex semantic web tasks [258].

Following the approach of [257], we initiated the interaction by providing prompt-based inputs. These inputs included the overall architecture of the KG (as outlined in Chapter 9), the schema prefixes used, and snippets of the Turtle code from all the graphs in the KG. After this initial training phase, we proceeded to generate the queries needed for this study.

10.3 Query Examples

10.3.1 Query 1: Price of a Product on a Specific Date and Location

One common type of query users perform in agricultural price index datasets is retrieving the price of a specific commodity on a given day in a particular location in Brazil.

To simulate this use case, we interacted with Chat GPT to design a query to retrieve the price of 'Café em grãos' (coffee beans). We provided the following instructions to Chat GPT:

> First, retrieve the dcat:Datasets entries that have 'Café em grãos' as the product type. For this, the query must first find the URI for 'Café em grãos,' which is given in the alm:productType attribute of the dataset. This URI is labeled as 'Café em grãos' in the APTO graph, which is also part of this KG. Once the URI is identified, the query can find datasets with alm:productType corresponding to 'Café em grãos'. After that, the query should filter the results for datasets with the sdo:location 'Minas Gerais'. However, the label 'Minas Gerais' is stored in the GeoNames ontology and is not directly part of the dataset attributes. The dataset attributes include the URI for the location, but the location name is stored in the GeoNames graph. In GeoNames, each location URI is defined as a gn:Feature, and the location name, such as 'Minas Gerais', as a gn:name. This query should ensure that we only retrieve observations for a specified date, for example '2023-06-01'.

ChatGPT's initial query was not fully functional due to two issues. First, it incorrectly used the language tag 'pt' instead of 'pt-br'. Second, it failed to retrieve the product type by its label because it did not account for the use of named graphs in the KG. To address this, we reminded ChatGPT to query the appropriate named graph, specifically graph:apto, when looking up product types by label. We also modified the query to declare a prefix for the named graphs at the beginning, improving readability and avoiding repetition – something the original query did not include.

Query Breakdown

1. **Identifying the Product Type**: The first step in the query was to ensure it correctly identified datasets that referenced coffee beans as a product. For that, the query first retrieves the product type by its label in Portuguese in the named graph:apto, as shown below:

```
GRAPH graph:apto {
   ?productType rdfs:label "Café em grãos"@pt-br .
}
```

2. **Linking the Product to Datasets**: Once the product type is identified, the query proceeds to find dcat:Datasets containing that product type:

```
          ?dataset a dcat:Dataset ;
        alm:productType ?productType ;
        sdo:location ?location ;
        alm:hasObservation ?observation .
```

3. **Retrieving Observation Data**: In the graph:data, prices are represented as blank nodes linked through the alm:hasObservation property. Each observation includes **date**, **price**, and **currency**. In this case, we retrieved data for a specific date:

```
      ?observation sdo:date ?date ;
           sdo:price ?price ;
           sdo:currency ?currency .
      FILTER (?date = "2023-06-01"^^xsd:date)
```

4. **Linking the Location to GeoNames**: Each dataset includes a reference to the location where the price was observed, represented as a URI in the `sdo:location` property. However, this URI alone is not human-readable, so the query utilizes the GeoNames Graph to extract the name of the location. The GeoNames URI is dereferenced to obtain the gn:name property, which contains the name of the location, such as ‘Minas Gerais’ in this case.

```
    ?location gn:name ?locationName .
    FILTER (str(?locationName) = "Minas Gerais")
```

Query Output

By combining these elements, the final output of the query includes the following fields:

```
SELECT ?locationName ?productType ?date ?price ?currency
WHERE {...}
```

This full query is accessible through the URI `https://data.pldn.nl/FilipiSoares/-/queries/Query-1-1/9`. A backup is also maintained on Zenodo [18]. The query results are shown in Table 10.1.

Table 10.1 Price of Coffee Beans in Minas Gerais on June 1, 2023

Location	Product Type	Date	Price	Currency
Minas Gerais	`http://aims.fao.org/aos/agrovoc/c_28379`	2023-06-01	893.82	BRL

10.3.2 Query 2: Time Series Data Visualization

We assume a user wants to see the evolution of the price for ‘Boi Gordo’ across the year 2022 based on the prices published by CEPEA. These prices are published on a daily but irregular basis, since CEPEA does not publish price index data on weekends and holidays. Then, the goal is to calculate the mean price for each month of 2022 and plot a chart with the results. We provided the following instructions to Chat GPT:

> Now, lets do another query. I want to show the mean monthly prices for ‘Boi gordo’, and this dataset dct:creator must be ‘Centro de Estudos Avançados em Economia Aplicada (CEPEA).’ CEPEA publishes data on a daily basis, but not regularly (it does not publish prices every day). Then, to calculate the month mean, this query has first to find all the prices of a given month, sum these prices, and then divide by the number of data points in that given month. After that, it should store this mean as a variable MonthMean. This should be done with all the data for the year of 2022. The results should show two columns: month/year (e.g, 03-2022) and MonthlyMean (with the mean prices).

The query provided by Chat GPT worked exactly as expected on the first attempt, with no modifications needed. This query is described below.

Query Breakdown

1. **Filtering by Dataset Creator and Product Type**: The first step of this query is to filter the dataset by the *creator*, which in this case is ‘Centro de Estudos Avançados em Economia Aplicada (Cepea).’

This ensures that only data published by Cepea is included in the query results. Additionally, it filters the datasets by the product type ‘Boi Gordo’, as shown below:

```
?dataset a dcat:Dataset ;
   dct:creator "Centro de Estudos Avançados em
   Economia Aplicada (Cepea)"^^xsd:string ;
   alm:productType <https://w3id.org/APTO#Boi_gordo> ;
   alm:hasObservation ?observation .
```

2. **Retrieving Daily Price Observations**: Each dataset in the KG contains daily price observations linked through the alm:hasObservation property. These observations include the *price* of Boi Gordo on a specific *date*, recorded using the properties sdo:price and sdo:date. The query filters the observations to include only those from 2022.

```
    ?observation sdo:date ?date ;
           sdo:price ?price .
```

3. **Grouping and Summing Prices by Month**: Once the daily observations are retrieved, the query groups the data by *month* and *year*. This grouping is done using the SPARQL functions `YEAR(?date)` and `MONTH(?date)`. The total price for each month is calculated by summing the prices of all the observations within that month using the `SUM(?price)` function.

```
SELECT (CONCAT(STR(MONTH(?date)), "-", STR(YEAR(?date))) AS
    ↪ ?MonthYear)
     ((SUM(?price)) AS ?MonthlyTotal)
WHERE {
  ?observation sdo:date ?date ;
           sdo:price ?price .
}
GROUP BY YEAR(?date) MONTH(?date)
```

4. **Counting the Number of Data Points**: Since the publication of prices by CEPEA is not regular (i.e., prices are not available for every single day), the query counts the number of price data points for each month using the `COUNT(?price)` function. This gives the number of observations for a given month, allowing us to divide the total sum of prices by the number of data points to calculate the *monthly mean*.

```
SELECT (COUNT(?price) AS ?MonthlyCount)
WHERE {
  ?observation sdo:date ?date ;
           sdo:price ?price .
}
GROUP BY YEAR(?date) MONTH(?date)
```

5. **Calculating the Monthly Mean Price**: The *monthly mean* is calculated by dividing the sum of the prices for the whole month by the number of observations. This is done using the formula:

$$\text{MonthlyMean} = \frac{\text{SUM(?price)}}{\text{COUNT(?price)}}$$

6. **Formatting the Output**: The final step in the query is to format the output. Instead of displaying the year and month as separate fields, the query concatenates them into a single string in the format MM-YYYY (e.g., 03-2023) using the `CONCAT` function. This provides a user-friendly way of viewing the results. The output displays only two columns: the month/year and the monthly mean price.

Query Output

The query result is shown in Fig. 10.1. This query and its results are accessible through the URI `https://data.pldn.nl/FilipiSoares/-/queries/Query-2/1`. A backup is also maintained on Zenodo [18].

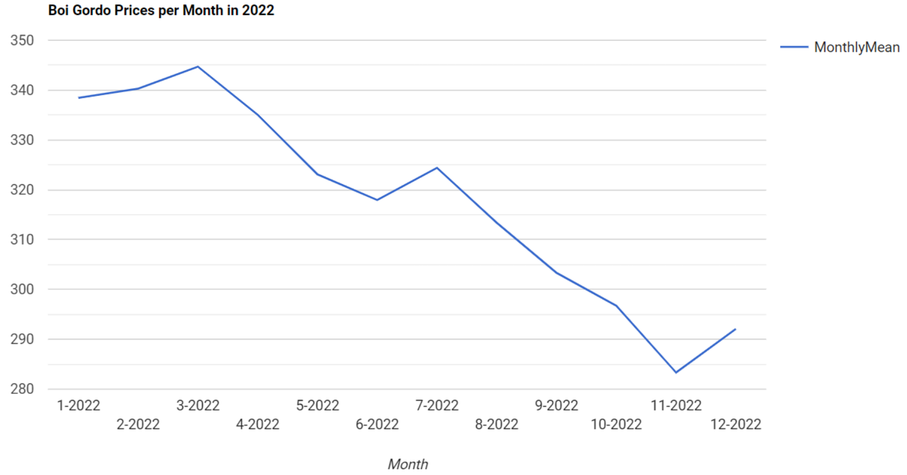


Figure 10.1 Query Results - Mean Price for ‘Boi Gordo’ in 2022.

10.3.3 Query 3: Comparing Prices from Two Organizations

The goal of this query was to compare the monthly prices of Cane Sugar (Açúcar Cristal) from CEPEA and CONAB, for the first semester of 2023 in the state of São Paulo. CEPEA publishes prices on a daily basis, whereas CONAB’s prices are typically monthly or weekly. The query aggregates the daily prices from CEPEA to calculate monthly averages and compares them to CONAB's monthly prices for the same time period.

Due to the complexity of this query, multiple interactions with the Chat GPT prompt were necessary. In these interactions, the following instructions have been given until it provided an appropriate query:

> In this query, we will compare the price indexes from two different institutions, for the same location. CONAB and CEPEA both publish prices for the state of São Paulo. We want to compare the monthly prices from both institutions for the same product: Cane sugar (Açúcar cristal). This comparison should be done for the first semester of 2023.
>
> CEPEA publishes daily prices, then the query should first calculate the mean price per month for CEPEA before comparing the prices with CONAB (which already publishes monthly prices).
>
> This query should also locate the datasets in the metadata graph and then the observations as specified.
>
> We are using a Virtuoso enviroment to run the SPARQL queries.

This query contains two subqueries, each designed to retrieve and calculate monthly prices for CEPEA and CONAB, respectively. The results from the two subqueries are then combined to compare the prices from both institutions.

Subquery 1: Calculating Monthly Mean Prices for Cepea

This first subquery retrieves daily prices for Cane Sugar published by CEPEA and aggregates them to produce monthly mean prices. It performs the following operations:

1. **Locating the Dataset**: The query first identifies datasets created by CEPEA using the dct:creator property. It further ensures that the dataset is for Cane Sugar by filtering on the alm:productType with the URI representing Cane Sugar (Açúcar Cristal). Additionally, the location is specified as São Paulo using the corresponding GeoNames URI.
2. **Retrieving Daily Prices**: Once the relevant dataset is located, the query retrieves the daily price (sdo:price) and the date (sdo:date) of the price observation:

```
  ?observation sdo:price ?price ;
               sdo:date ?date .
```

3. **Filtering by Date (First Semester of 2023)**: The query filters the observations to include only those from the first six months of 2023:

```
  FILTER (YEAR(?date) = 2023 && MONTH(?date) <= 6)
```

This ensures that only data from January to June of 2023 is considered.

4. **Grouping by Month and Year**: The query uses the `GROUP BY` clause to group the daily prices by month and year:

```
  GROUP BY YEAR(?date) MONTH(?date)
```

This groups all the daily prices for each month into a single group.

5. **Calculating the Monthly Mean Price**: For each group of daily prices (one group for each month), the query calculates the monthly mean price by dividing the sum of daily prices by the number of daily observations:

```
  (SUM(?price) /COUNT(?price)) AS ?cepeaMonthlyMean
```

This produces and stores the monthly mean price for CEPEA for each month.

Subquery 2: Retrieving Monthly Prices for CONAB

The second subquery retrieves the monthly prices for Cane Sugar published by CONAB. This subquery works similarly to the first but assumes that CONAB publishes data on a monthly basis rather than daily. It works as follows:

1. **Locating the Dataset**: The query locates datasets published by CONAB using the `dct:creator` property; It filters for datasets related to Cane Sugar (Açúcar Cristal) and ensures that the location is São Paulo, similar to the first subquery.
2. **Retrieving Monthly Prices**: The query retrieves the price (`sdo:price`) and date (`sdo:date`) of the observation. Since CONAB is assumed to publish monthly data, the price retrieved corresponds to a full month:

```
  ?observation sdo:price ?price ;
              sdo:date ?date .
```

3. **Filtering by Date (First Semester of 2023)**: As with CEPEA, the query filters CONAB's data to include only observations from the first six months of 2023:

```
  FILTER (YEAR(?date) = 2023 && MONTH(?date) <= 6)
```

4. **Grouping by Month and Year**: The query groups the prices by month and year, similar to CEPEA's subquery:

```
  GROUP BY YEAR(?date) MONTH(?date)
```

5. **Calculating the Average Monthly Price**: For each month, the query calculates the average price for CONAB using the `AVG(?price)` function:

```
  AVG(?price) AS ?conabPrice
```

6. **Handling Missing Data**: The `OPTIONAL` keyword in the subquery for CONAB ensures that if there are months in which CONAB data is missing, it should print `NULL` values.

Combining the Results and Final Output

Once the monthly mean prices for CEPEA and CONAB are calculated in their respective subqueries, while the outer query combines the results to provide a side-by-side comparison of prices for each month.

The outer query groups the results by month using the `?month` variable, which is derived by concatenating the month and year:

```
(CONCAT(STR(MONTH(?date)), "-", STR(YEAR(?date))) AS ?month)
```

The final output shows the monthly mean price for CEPEA (calculated from daily data) and the average monthly price for CONAB:

```
  SELECT ?month (AVG(?cepeaMonthlyMean) AS ?CepeaPrice) (AVG(
      ↪ ?conabPrice) AS ?ConabPrice)
```

The query orders the results chronologically by month using the `ORDER BY ?month` clause. The results were used to plot the chart shown in Fig. 10.2, where it is possible to compare the prices from both institutions. This query and its results are accessible through the URI `https://data.pldn.nl/FilipiSoares/-/queries/Query-3/1`. A backup is also maintained on Zenodo [18].

Changes Introduced in the Second Query for Adjusting Prices Per Kg

Fig. 10.2 shows that cane sugar prices reported by CEPEA were higher than those from CONAB during the first semester of 2023. However, this comparison requires caution: CEPEA's prices are based on a 50kg bag of sugar (as defined in `sdo:referenceQuantity`), while CONAB's prices are based on a 30kg bag, so this comparison is not directly proportional. This difference can be addressed by adjusting the SPARQL query to account for the different reference quantities. To update this query, we provided Chat GPT with the following instructions:

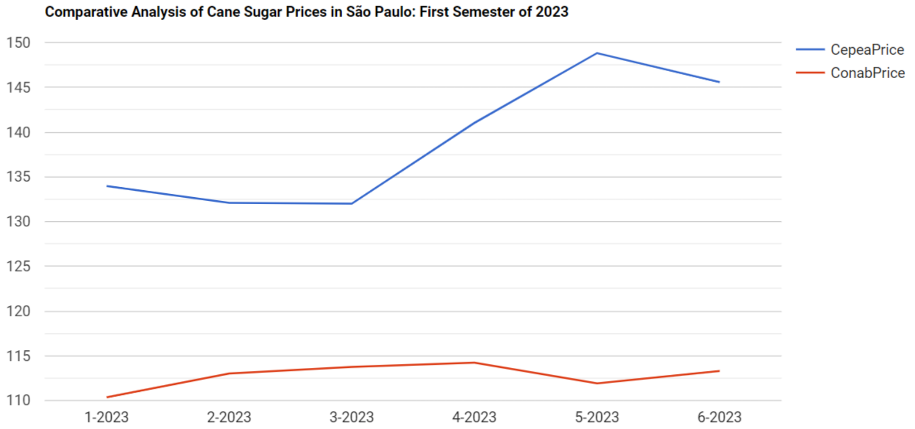


Figure 10.2 Query Results - Prices from CEPEA and CONAB for Cane Sugar for the First Semester of 2023.

> CONAB provides prices for a 30kg bag of sugar, while CEPEA reports prices for a 50kg bag. As a result, CEPEA's prices appear significantly higher, but this comparison is misleading due to the difference in product quantities. To make the comparison fair, the query should be adapted to automatically calculate and adjust all prices based on a 30kg bag.

The key change in this second version is the adjustment of CEPEA's prices to match the 30kg bag weight used by CONAB, to ensure that the price comparison reflects the same product quantity. The main differences are as follows:

1. **Price Adjustment for CEPEA**: In the second query, CEPEA's prices, which are originally reported for 50kg bags, are adjusted to a 30kg basis. This is done using the `BIND` function:

   ```
   BIND(?price * 0.6 AS ?adjustedPrice)
   ```

   The conversion factor of 0.6 is applied because

   $$\frac{30}{50} = 0.6$$

   meaning that to standardize CEPEA's prices to 30kg, we multiply each price by 0.6.

2. **Renaming Variables for Clarity**: After adjusting CEPEA's prices, the calculated price is stored in the variable `?adjustedCepeaPrice`. This replaces `?cepeaMonthlyMean` used in the original query:

   ```
   (SUM(?adjustedPrice) /COUNT(?adjustedPrice)) AS
       ↪ ?adjustedCepeaPrice
   ```

This change ensures that the final result shows CEPEA's monthly prices for 30kg bags under the alias `?CepeaPrice30kg`, making the comparison with CONAB's prices more meaningful.

3. **No Change to CONAB Prices**: CONAB's prices are already reported for 30kg bags, so no adjustments are necessary for this part of the query. The variable `?conabPrice` remains unchanged, and the calculation for CONAB's monthly prices is performed as it was in the first version of the query.

Final Output Adjustment

The initial price comparison between CEPEA and CONAB appeared to show that CEPEA's sugar prices were higher. However, this analysis was based on different bag sizes: CEPEA reports prices for 50kg bags, while CONAB uses 30kg bags. After normalizing CEPEA's prices to a 30kg equivalent, a more accurate comparison became possible. As shown in Fig. 10.3, the adjusted analysis reveals that CEPEA's sugar prices are actually lower than CONAB's – a reversal of the initial impression shown in Fig. 10.2.

This discrepancy can be explained by the economic principle of economies of scale. In Brazil, as in many markets, purchasing a larger quantity of a product – such as a 50kg bag of sugar – often results in a lower price per unit of weight due to bulk discounts. Sellers reduce the price per kilogram as the order size increases, which reflects operational savings in packaging, logistics, and transaction costs. Therefore, CEPEA's lower per-kilo price, once adjusted for bag size, aligns with this market dynamic.

This query and its results are accessible via the URI `https://data.pldn.nl/FilipiSoares/-/queries/Query-3/2`, and a backup is maintained on Zenodo [18].

10.3.4 Query 4: Comparing Prices Between Two Locations

As done in [248], some price index data analysis involves comparing prices from two or more regions in Brazil. In this KG, this type of analysis can also be simulated with a SPARQL query. We used CONAB data as starting point. Suppose someone wants to compare the prices for Soybean between Paraná and Mato Grosso, the two most important soybean producers in Brazil, over the year 2023. For that we provided the following instructions to Chat GPT:

> To compare CONAB prices for the locations Paraná and Mato Grosso, this query should first retrieve datasets where CONAB is the creator and the location is either Paraná or Mato Grosso. Then, it should extract the relevant observations (price, date, and currency) for both datasets, storing them in separate

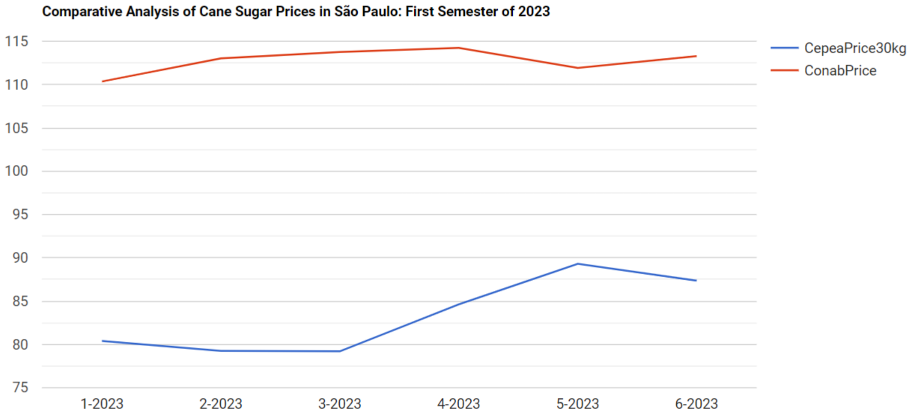


Figure 10.3 Harmonized query results showing prices from CEPEA and CONAB for cane sugar for the first semester of 2023.

variables. Next, it should generate a column named 'Date' that contains only the month and year (excluding the day). Additionally, it should create two columns: SojaParanaPrice for Paraná prices and SojaMatoGrossoPrice for Mato Grosso prices, omitting the currency in this case.

On the first attempt, the query failed due to an issue with handling the date format. The dates were provided in the YYYY-MM-DD format, and the query needed to truncate them to just the year and month to properly match observations between Paraná and Mato Grosso.

Additional instructions were needed after the next version of the query failed to filter by product type. We also shared snippets of the KG to improve query accuracy. After several iterations, a functional query was generated.

Query Breakdown

1. **Selecting the Datasets**: The query begins by identifying the relevant datasets for Paraná and Mato Grosso using dct:creator, alm:productType, and sdo:location. The `OPTIONAL` clause ensures the query does not fail when data for Mato Grosso is missing for a given month (We observed that the publication of prices for Mato Grosso was irregular in some years).
2. **Retrieving Observations**: After identifying the datasets, the query retrieves the soybean price and the date of the observations for each state.

   For Paraná (always available):

```
  ?observationParana sdo:price ?paranaPrice ;
```

```
                   sdo:date ?dateParana .
```

For Mato Grosso (if it exists):

```
   OPTIONAL {
     ?observationMatoGrosso sdo:price ?matoGrossoPrice ;
                          sdo:date ?dateMatoGrosso .
   }
```

3. **Ensuring Matching Dates for Both States**: To ensure that the prices for Mato Grosso and Paraná are compared on a monthly basis, a `FILTER` is used to align the dates.
4. **Extracting Month and Year**: The query extracts the year and month (in YYYY-MM format) from the date of the observations from Paraná:

```
   BIND(SUBSTR(STR(?dateParana), 1, 7) AS ?month)
```

   This ensures that the monthly average prices concern the same time period.
5. **Filtering Data for 2023**: To select the year 2023, the query includes a `FILTER`:

```
   FILTER (STRSTARTS(STR(?dateParana), "2023"))
```

6. **Calculating Monthly Averages**: For Paraná, the query averages the prices for each month:

```
   (AVG(?paranaPrice) AS ?avgParanaPrice)
```

   For Mato Grosso, if data exists for the same month, the average price is calculated similarly:

```
   (AVG(?matoGrossoPrice) AS ?avgMatoGrossoPrice)
```

Final Output

The query groups results by month and year, displaying them in chronological order. The results are presented in a table format, from which the chart in Fig. 10.4 was generated. The chart indicates similar fluctuations in both regions for soybean prices throughout the year. Additionally, it shows that soybean prices in Paraná remained consistently lower than those in Mato Grosso over the entire year.

This query and its results are accessible via the URI `https://data.pldn.nl/FilipiSoares/-/queries/Query-4/1`, and a backup is maintained on Zenodo [18].

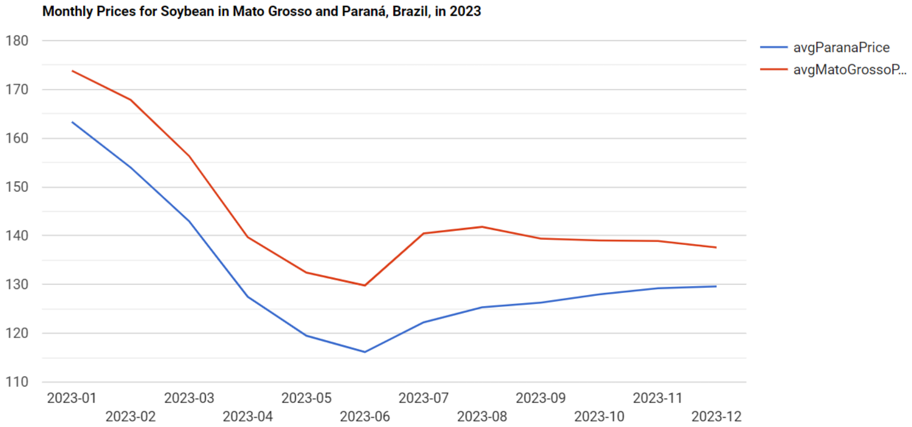


Figure 10.4 Prices for Soybean in Paraná and Mato Grosso in 2023.

10.4 Benefits and Limitations

In this study, we explored an approach for integrating and comparing data from various datasets using Semantic Web technologies and SPARQL queries. The methodology centered on querying data from multiple sources, specifically focusing on agricultural price indexes from different institutions. By using RDF as the underlying data structure and SPARQL as the query language, we were able to access, query, and compare datasets in a highly flexible and structured way. This approach provided significant benefits for data interoperability, analysis, and comparison, but it also revealed some limitations related to data structure and querying processes. Furthermore, using ChatGPT to assist in generating and refining SPARQL queries added value to this approach but also introduced certain challenges.

10.4.1 Using ChatGPT for SPARQL Query Generation

ChatGPT offered substantial support throughout the query development process. By generating and refining SPARQL queries, it reduced the time required to construct complex queries and helped overcome many of the technical hurdles associated with querying deeply nested RDF structures.

ChatGPT was able to suggest efficient query structures that combined filtering, grouping, and aggregation operations. For instance, it helped automate complex tasks like aligning daily and monthly price data across datasets. When faced with RDF graphs involving blank nodes or deeply nested structures (such as observations within datasets), ChatGPT efficiently handled these complexities by suggesting multi-step query strategies, reducing manual trial-and-error efforts. ChatGPT's ability to quickly generate SPARQL code based on user input allowed for rapid prototyping and experimentation. This was particularly useful when iterating on queries to handle different date formats, geographic relationships, or data alignment challenges.

However, while ChatGPT provided valuable assistance, there were also limitations in its ability to fully handle the intricacies of SPARQL and RDF querying. Although ChatGPT could generate valid SPARQL queries, it occasionally struggled with the broader context of the datasets. For example, queries required adjustment when additional data structures (such as blank nodes or specific ontology terms) were introduced. ChatGPT sometimes missed these intricacies, requiring manual refinement and deeper exploration of the data structure. In cases where datasets had inconsistent structures or required specific filters (e.g., price adjustments, product-specific handling), ChatGPT's query suggestions sometimes needed fine-tuning. For example, when querying data for both daily and monthly price comparisons, additional manual steps were necessary to ensure that temporal alignment was accurate. In addition, the following limitations were perceived:

- **Context Sensitivity Across Named Graphs**: SPARQL queries that span multiple named graphs (e.g., metadata, data, and GeoNames graphs) require careful handling. ChatGPT initially did not correctly assign parts of the query to the appropriate graphs, which led to no results being returned. This indicates that while ChatGPT understands the RDF and SPARQL syntax, it can struggle with nuanced context-dependent requirements like federated graph querying or data distributed across named graphs.
- **Lack of State Memory Across Query Components**: When composing multi-step queries – such as retrieving a URI using a label, then using that URI in subsequent filters – ChatGPT sometimes failed to carry over intermediate variables correctly. This reflects a limitation in managing multi-step logical flow within SPARQL, especially when it involves dereferencing, binding, and using dynamic values from earlier steps.
- **Misinterpretation of Language Tags**: Language tags (e.g., @pt-br vs @pt) significantly affect RDF label matching. ChatGPT needed manual correction when the label "Café em grãos"@pt-br was not found, demonstrating that default assumptions about RDF literals and language tags can lead to incorrect results unless explicitly handled.
- **Date Handling Complexity**: Extracting month and year from full xsd:date values and grouping appropriately posed repeated challenges. Some queries failed due to missing GROUP BY components for projected variables. These are syntactical oversights common in complex SPARQL query construction.

- **Verbose Trial-and-Error Iteration:** The process of developing SPARQL queries was highly iterative, involving extensive back-and-forth refinement. For the four queries discussed in this chapter, a total of 147 interactions with ChatGPT were recorded, comprising both user inputs and ChatGPT responses. This volume reflects the complexity of aligning multiple graphs, adjusting query logic, resolving syntax issues, and validating results. While ChatGPT provided valuable support, the iterative nature highlights the need for human oversight and fine-tuning in real-world Semantic Web querying.

While ChatGPT could assist in query generation, it was less effective in debugging errors related to specific SPARQL engines or syntax peculiarities. For instance, some queries that appeared correct in theory yielded execution errors when performed by Virtuoso. This required manual troubleshooting beyond the scope of ChatGPT's support.

10.4.2 Overall Approach

The use of RDF and SPARQL allowed integrating datasets from different institutions, with varying formats and frequencies of data publication. RDF provides a universal data model that supports the representation of any data structure, while SPARQL enables querying across different datasets that share common vocabularies. This flexibility was especially useful to compare datasets with data coming from different states (i.e., Paraná and Mato Grosso) for the same product, or even when the datasets were published on different scales (daily vs. monthly).

One of the key advantages of using SPARQL to query RDF data is the ability to work with heterogeneous data. In our queries, we dealt with different locations, products, and timeframes, yet SPARQL's filtering and grouping capabilities allowed us to align and aggregate this data. This approach also allowed us to perform automated calculations within queries, which eliminated the need for manual calculations.

A key limitation of this case study was the inability to import the complete geographic hierarchy due to API request limits. The GeoNames API restricts requests to 1,000 credits per hour, making it difficult to retrieve full geographic hierarchies (e.g., parent/child relationships) for larger datasets, such as the entire country of Brazil. This restriction limited our ability to conduct more in-depth spatial analysis. To overcome this, a paid subscription to the GeoNames API would be required to increase the number of allowable requests and retrieve more extensive geographic data.

While RDF provides great flexibility in representing data, querying deeply nested structures or blank nodes increased the complexity of the SPARQL queries. This was particularly evident when querying product observations, which required multiple levels of indirection. Managing these complexities required additional effort in structuring and debugging queries, sometimes making the process less efficient than working with simpler relational databases.

Another challenge was dealing with datasets that had different time granularities (e.g., daily vs. monthly). Aligning dates from different datasets required significant use of functions like `SUBSTR` to extract year and month from date fields. While this approach worked, it introduced additional complexity when comparing datasets. More sophisticated temporal functions or data alignment strategies could be beneficial to reduce this complexity.

10.5 FAIR Assessment

By querying the KG we confirmed that it successfully answers the queries it was designed for, demonstrating that it meets its intended objectives. However, an additional evaluation dimension is the KG's level of FAIRness, given that the semantic interoperability framework presented in this dissertation is based on the FAIR Data Principles. Consequently, C4AI-KGAP, which was developed using this framework, should comply with the FAIR Data Principles and exhibit higher FAIRness compared to the original datasets from CEPEA, CONAB, and Ipea, from which it derives its data.

Assessing the FAIRness of datasets is a complex task involving multi-level analysis of the elements that contribute to FAIR compliance. This assessment can be either automated or manual, each with its own advantages and disadvantages. Automated assessments tend to focus on standardized schemas (e.g., Dublin Core, schema.org, DCAT) and identifier services such as PURL and DOI, which are widely known and implemented. However, automated tools may overlook domain-specific resources, potentially leading to lower scores if the datasets do not fall within the predefined rules of the tool [259]. In contrast, manual assessments (typically conducted by specialists such as Data Stewards) are more flexible, as they can evaluate domain-specific elements. However, this flexibility introduces potential bias, as personal judgment can affect the evaluation outcomes [259].

To balance these limitations, we adopted a hybrid approach in this case study. Initially, we used an automated tool for the initial FAIR assessment, followed by a qualitative review of the results.

The work of [259] provided a comprehensive review of existing tools and metrics for assessing FAIRness, highlighting their limitations and gaps. Among the 20 tools tested, we selected the F-UJI Automated FAIR Data Assessment Tool for our evaluation, because it is domain-independent, performs automated analysis [259], and is an output of the FAIRsFAIR project, which also inspired the requirements for the FAIR metadata schemas conceptual model presented in Chapter 3. Additionally, F-UJI has been previously used to assess the FAIRness of agricultural datasets [260]. The tool assesses research data objects at the dataset level based on FAIRsFAIR Data Object Assessment Metrics [261]. The process simply involves providing a URL to the data object and running the assessment.

F-UJI currently applies FAIRsFAIR (FsF) metrics to assess the FAIRness of data objects according to the FAIR principles. These metrics include specific checks, such as the presence of unique identifiers (Findability), access protocol availability (Accessibility), use of standard metadata and vocabularies (Interoperability), and clear licensing and provenance information (Reusability). F-UJI applies these metrics to evaluate datasets and provides a score indicating how well the data adheres to the FAIR data principles, offering detailed feedback on areas of improvement.

We started by assessing the FAIRness of the original datasets from CEPEA, Ipea and CONAB, and then we assessed the FAIRness of the C4AI-KGAP itself to assess the improvements in FAIR compliance compared to the original datasets [262].

10.5.1 FAIRness of CEPEA, IPEA, and CONAB Datasets

The FAIR assessment of CEPEA's dataset was conducted using the public URL[1], where data can be downloaded in Excel format. The resulting score is shown in Fig. 10.5.

For the Findability principle, the dataset had a globally unique identifier, but it was not persistent. The descriptive metadata provided was minimal, consisting mostly of generic Dublin Core fields embedded in HTML. Key metadata elements essential for citation were missing.

Regarding the Accessibility principle, no direct access information was provided in the metadata, and the metadata lacked a direct link to the data. However, the metadata was accessible via standardized communication protocols, partially meeting accessibility requirements.

For Interoperability, the dataset scored no points, primarily because it did not use any semantic resources like ontologies or metadata schemas. No formal knowledge representation languages (e.g., JSON-LD, RDFa) were used either, and no links were established between the dataset and related entities.

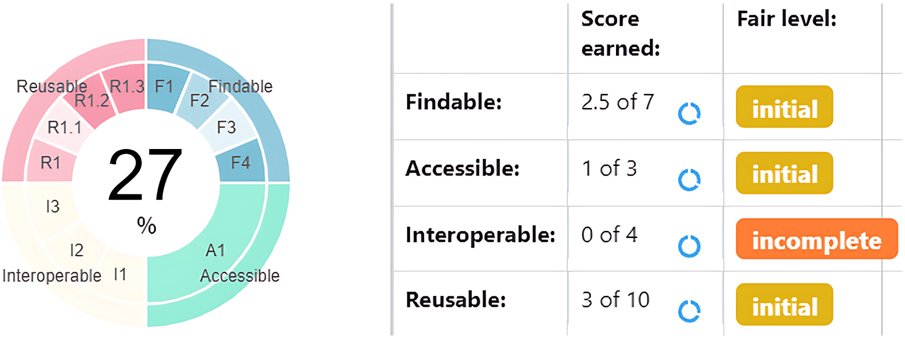
As for Reusability, minimal content information was provided. Although the dataset was labeled with a CC BY-NC 4.0 license, the automated assessment tool could not detect it. Provenance information was sparse and related more to the website creator than the dataset creator. While Dublin Core metadata was used in the HTML embedding, no com-

---

[1] `https://www.cepea.esalq.usp.br/br/consultas-ao-banco-de-dados-do-site.aspx`

munity schema was employed to describe the data comprehensively. Additionally, only the Excel file format was available.

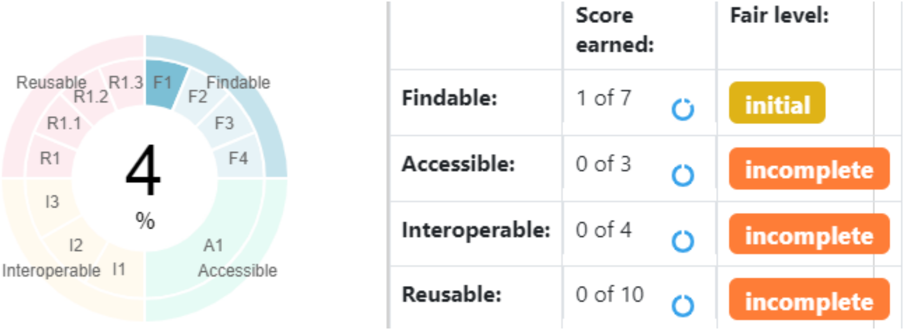
In conclusion, the CEPEA dataset performed poorly across the FAIR metrics, particularly in Interoperability and Reusability. Improvements, such as the adoption of persistent identifiers, semantic metadata, and community standards, would significantly enhance its FAIRness.



	Score earned:	Fair level:
Findable:	2.5 of 7	initial
Accessible:	1 of 3	initial
Interoperable:	0 of 4	incomplete
Reusable:	3 of 10	initial

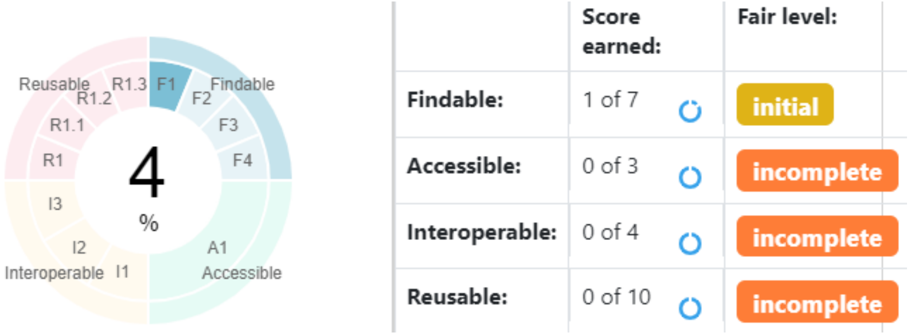
Figure 10.5 Assessment Results for the Cepea Dataset.

For Ipea and CONAB, the FAIRness scores were even lower, at just 4%, as shown in Fig. 10.6 and Fig. 10.7. The only FAIR principle they fulfilled was assignment of a unique identifier to their datasets. Although both Ipea and CONAB use metadata to describe their data, they do not follow any standardized formats or structure that allows machines (such as the F-UJI engine) to process the data. In Ipea's case, the dataset is embedded within a complex table structure in the JavaScript code. Similarly, CONAB's dataset does not comply with any community standard and fails to meet any other FAIR principles.



	Score earned:	Fair level:
Findable:	1 of 7	initial
Accessible:	0 of 3	incomplete
Interoperable:	0 of 4	incomplete
Reusable:	0 of 10	incomplete

Figure 10.6 Assessment Results for the Ipea Dataset.

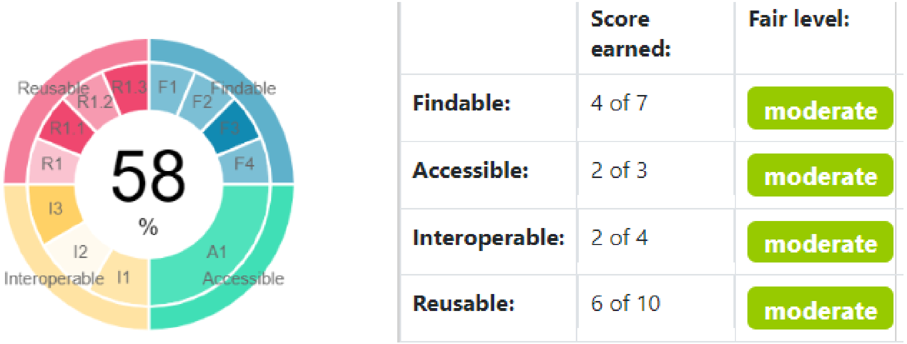


	Score earned:	Fair level:
Findable:	1 of 7	initial
Accessible:	0 of 3	incomplete
Interoperable:	0 of 4	incomplete
Reusable:	0 of 10	incomplete

Figure 10.7 Assessment Results for the CONAB Dataset.

10.5.2 FAIRness of the C4AI-KGAP

As shown in Fig. 10.8, the C4AI-KGAP performed significantly better than the original datasets in terms of FAIRness, although it achieved only a moderate score of 58%. As with the previous datasets, we carefully reviewed the results of the assessment to identify areas for improvement.



	Score earned:	Fair level:
Findable:	4 of 7	moderate
Accessible:	2 of 3	moderate
Interoperable:	2 of 4	moderate
Reusable:	6 of 10	moderate

Figure 10.8 Assessment Results for the C4AI-KGAP.

Starting with the Findability principle, the F-UJI engine flagged that the data identifiers are not persistent. This is accurate, as we used GitHub hashes as identifiers, which are not guaranteed to be persistent, i.e., if a file is deleted or moved, the identifier becomes invalid. We are aware of this limitation, and as we move forward to implement C4AI-KGAP for real-world use, this issue will be addressed.

Another issue related to Findability was the engine's failure to parse the RDF data, even though the entire KG is modeled in RDF/Turtle. Upon inspection, we found that the engine attempted to access the RDF serialization from the root URL of the data repository, but the graphs in Turtle format are stored in subdirectories. The engine should have been able to access them, so this may indicate a problem with how the tool interprets data structures.

The engine also noted the absence of core descriptive metadata elements, such as keywords. Although we used alm:productType, alm:productGroup, and alm:theme to describe the dataset content, we agree that these fields should not replace traditional keywords. While they serve a similar purpose, we acknowledge that including conventional keywords would make it better understandable to broader communities. This will be improved in future developments.

For Interoperability, the engine incorrectly reported that the dataset was not accessible through a SPARQL endpoint, even though one is available. This is likely due to a tool limitation in locating the SPARQL endpoint.

Another issue flagged under Interoperability was the absence of known vocabulary namespaces listed in the Linked Open Data (LOD) registry. This is inaccurate, as we used well-established vocabularies such as schema.org, Dublin Core, and DCAT. The only exception is Almes Core, which is not yet registered in LOD. This issue may stem from the tool's failure to parse the vocabularies correctly.

Under Accessibility, the tool reported that no information on access restrictions or rights was provided, and that access information was not machine-readable. However, the KG uses Creative Commons licenses, which are machine-readable. The tool's failure to detect this may result from how PLDN displays accessibility information, even though it is present.

Lastly, for Reusability, the tool pointed out missing file size metadata and the absence of measured variables. These elements will be added in the next version of the KG. Additionally, while provenance information was provided using Dublin Core, the tool recommended using the PROV-O ontology. While we agree that PROV-O is an appropriate standard, we believe that Dublin Core and DCAT also adequately describe provenance. The tool's preference for PROV-O highlights a limitation in its scope.

10.6 Final Considerations

Overall, the integration of RDF and SPARQL, coupled with the support from ChatGPT, provided a powerful framework for querying and analyzing agricultural price data. While ChatGPT proved useful in generating and refining SPARQL queries, especially for complex tasks such as filtering and aggregating data, there were limitations in terms of contextual understanding, error debugging, and handling large-scale datasets. The flexibility of RDF, combined with SPARQL's query capabilities, enabled comprehensive data comparisons, though at the cost of additional complexity when handling nested structures, temporal alignment, and large data volumes.

Further advancements in AI-assisted query generation could potentially mitigate some of these limitations, particularly in terms of query optimization and automatic handling of complex RDF structures. Nonetheless, the combination of Semantic Web technologies and AI-driven support offers promising potential for enhancing data integration and analysis in fields such as agriculture.

In addition, the FAIR assessment showed a clear improvement in the C4AI-KGAP compared to the original datasets it integrated. However, the assessment also highlighted areas for enhancement, which will be considered in future versions of the KG. Additionally, we observed limitations in the F-UJI engine's ability to navigate the complex structure of the C4AI-KGAP, resulting in some inaccurate findings where FAIR-related elements were actually present.

Part III

Case Study II: Harmonization and Integration of Plant-Pollinator Interactions Datasets

Introduction to Part III

Agricultural biodiversity has been chosen as the second domain to test the semantic interoperability framework presented in Chapter 2. The scope of this case study is plant-pollinator interactions, which form a complex network crucial for ecosystem functionality and agricultural productivity. Pollinators, such as bees, butterflies, and other animals, play a vital role in the reproduction of many plants, including numerous crops. Approximately 75% of global food crops depend on animal pollination to some extent, making it an indispensable ecosystem service for food security [263]. Effective management of pollinator populations can significantly enhance crop yields and quality. Additionally, the decline in pollinator populations poses a threat to agricultural sustainability and biodiversity, underscoring the need for comprehensive data and research in this domain [264].

Despite the essential role of pollination in ecosystems and agriculture, data on this process is often scattered or incomplete, limiting conservation efforts. Managing biodiversity data poses challenges like taxonomic accuracy, with added complexity in representing biotic interactions such as pollination. The Darwin Core standard lacks the necessary properties for handling these interactions, underscoring the need for FAIR vocabularies. To address this, the Research Data Alliance (RDA) Improving Global Agricultural Data Community of Practice (IGAD CoP) has partnered with key stakeholders to enhance data interoperability and support pollinator conservation efforts [78].

CHAPTER 11

Semantic Interoperability in The WorldFAIR Project

Chapter Overview

This chapter explores the needs for semantic interoperability in the WorldFAIR Project's Agricultural Biodiversity Case Study, focusing on plant-pollinator interactions data. Drawing from collaborative efforts led by RDA IGAD CoP and partners like Embrapa, Meise Botanic Garden, and HiveTracks, the study aimed to align diverse datasets to facilitate reuse. To achieve this, the project deployed templates based on domain metadata schemas and mapped original datasets to these templates. Data from six pilot projects, representing different data collection methods, regions, and languages, were transformed into interoperable formats, applying domain-specific metadata schemas (e.g., DwC, PPI) and controlled vocabularies, including RO and AGROVOC. The chapter highlights challenges in semantic annotation and FAIR assessments, addressing human errors in taxonomy through AI-based bots to detect typos and standardize scientific names. Through GloBI's federated search platform, the project demonstrated the benefits of FAIRified data, including the automated generation of and data reviews.

11.1 Introduction

We actively participated in the WorldFAIR Agricultural Biodiversity project, contributing to multiple stages of its development and execution. This direct involvement provides the basis for using the project as a case study in this dissertation. Throughout this chapter, we highlight our specific contributions to the project, focusing on how the proposed semantic interoperability framework was applied in practice.

The core research team for this project included Débora P. Drucker (Embrapa), José A. Salim (Unicamp), Jorrit Poelen (GloBI), and Filipi M. Soares (the author of this dissertation). Each pilot initiative involved close collaboration with assigned researchers, who supported the work by sharing data, discussing institutional data policies, and addressing issues related to data access and reuse.

Given our participation, this chapter analyzes the project through the lens of the semantic interoperability framework developed in this research. We map each stage of the project to corresponding components of the framework, drawing from official project documentation and deliverables published from 2022 to 2024 [78, 79, 80, 81, 265].

11.2 Goal Definition

This section reports on how the process of Goal Definition (Section 2.3) was performed in this case study. This case study was part of the WorldFAIR project and was led by the RDA Interest Group on Agricultural Data (IGAD) Community of Practice (CoP), with coordination by Embrapa. The project objective was to improve the FAIRness of plant-pollinator interaction data by focusing on semantic interoperability among diverse institutions and their integration with the Global Biotic Interactions (GloBI) platform.

The overarching aim of this initiative was to harmonize access to plant-pollinator interaction datasets across global institutions in order to:

- Promote the adoption of FAIR data standards for documenting plant-pollinator interactions.
- Enhance semantic interoperability between institutions and platforms (e.g., GloBI) to support data discovery.
- Develop shared metadata templates, ontologies, and controlled vocabularies for data harmonization.
- Foster collaboration between organizations from different countries and contexts on issues of data sharing and infrastructure alignment.
- Contribute to international recommendations and best practices for agricultural biodiversity data stewardship.

The datasets involved in this case study include observational data on flower visitation, crop pollination services, species interaction networks, and pollinator behavior collected via field studies, literature reviews, and citizen science tools. These datasets were previously siloed in different formats, languages, and data models. Harmonizing their structure and semantics enables broader reuse of this data to support ecological research, conservation policy, agricultural decision-making, and cross-disciplinary analysis of pollination services – an essential ecosystem function tied directly to food production and biodiversity.

11.2.1 Selected Datasets

Six pilot projects were selected to reflect a wide range of data collection practices, geographic regions, and scientific objectives. Each pilot was supported by one or more institutions responsible for the data:

1. **Facultad de Agronomía, Universidad de Buenos Aires (UBA)** – Provided five datasets from field surveys in the Argentine Pampas, including over 120 plant-pollinator interaction networks [79].
2. **Brazilian Network on Plant-Pollinator Interactions - REBIPP (Brazil)** – Contributed six datasets on floral visitation, pollination effectiveness, and interaction networks involving native and crop plants across multiple Brazilian biomes [79].
3. **KALRO (Kenya)** – Compiled 1,030 plant-pollinator interaction records from African sources such as research articles, manuals, and databases [79].
4. **HiveTracks (USA)** – Supplied crowdsourced data from its mobile app used by beekeepers globally to track hive management, local flora, and pollinator activity [79].
5. **USDA** – Developed structured data tables to represent pollination-related interactions in the USDA NRCS PLANTS database, based on North American biodiversity literature [79].
6. **Universidad CES & SIB Colombia** – Provided datasets from systematic field surveys documenting floral visitors and pollinator interactions in Colombian ecosystems [79].

The pilot institutions provided both domain knowledge and original datasets. Embrapa led the coordination of the case study and data modeling efforts. Meise Botanic Garden contributed its expertise in biodiversity data digitization and standardization. GloBI provided technical infrastructure for indexing and exposing the harmonized datasets. The RDA IGAD CoP ensured alignment with broader agricultural data initiatives and helped promote international collaboration across biodiversity, agricultural, and data science communities [78].

11.3 Data Preparation

This section describes how the Data Preparation process (Section 2.4) was carried out. Due to the diversity of the datasets involved – each originating from different institutions, regions, and data collection protocols – a common data template was essential to ensure consistency and facilitate integration. As in Case Study 1 (see Section 4.3), defining the structure and scope of the data was a critical first step to determine which data fields would be included and which would be excluded from harmonization efforts.

To that end, the REBIPP template was adopted by the project team and participants from the pilots as the shared reference model for standardizing data and metadata across all pilot projects. Developed to support the representation of plant-pollinator interactions, this template enabled the mapping of heterogeneous data structures into a single, harmonized format, thereby enhancing semantic interoperability between the project partners [266].

Many of the original datasets contained additional columns beyond the scope of the REBIPP model. These were removed to align with the template. The remaining fields were then mapped to the REBIPP structure, which incorporates elements from the following community metadata schemas:

- Ecological Metadata Language (EML) [267] — used for capturing provenance, descriptive, and administrative metadata.
- Plant-Pollinator Interactions Vocabulary (PPI) [77] — used for the overall data structure, which is organized around three main classes: ppi:Plant, ppi:Interaction, and ppi:Animal [79].
- DwC [47] and PPI — used for labeling individual data columns within the above classes.

To ensure uniformity across datasets, data types were defined in accordance with DwC and PPI guidelines, covering elements such as taxonomic names, dates, geographic coordinates, and interaction types.

Once harmonized, all datasets were published in tabular format (.xlsx) and made publicly available on Zenodo. Each dataset is versioned and assigned a DOI for traceability and citation. The full list of records includes: [20, 21, 22, 23, 24, 25, 26, 27, 28, 29].

11.4 Data Semantic Annotation

This section discusses the Data Semantic Annotation process of the semantic interoperability framework (Section 2.5).

11.4.1 Identification of Domain Metadata Schemas

To ensure consistency in plant-pollinator interaction data description, the metadata schemas used in the templates mentioned in Section 11.3 were selected because they cover key elements such as species, location, observation time, and interactions. The DwC schema [47] was chosen due to its widespread use in biodiversity informatics. DwC supports structured descriptions of biodiversity data, capturing essential attributes like taxonomy, geolocation, and observation details. However, its application to species interactions presents challenges due to ambiguous interpretations of its elements. For example, while the term `dwc:associatedTaxa` is commonly used for interactions, the `dwc:ResourceRelationship` class offers a better alternative for representing interactions at the occurrence level [77, 268, 269, 270].

To address these challenges, Salim et al. [77] defined the PPI vocabulary to extend DwC for representing plat-pollinator interactions. The PPI conceptual model, shown in Fig. 11.1, places the interaction event (gray box) at the center, modeled as an instance of the `dwc:Event` class, reflecting the fact that interactions occur at specific times and places.

The core of the schema links two `dwc:Occurrence` records (blue boxes), representing the observed presence of a plant and a pollinator, to the `dwc:Event` through a shared `dwc:eventID`. These occurrence records capture taxonomic identification and metadata about each organism involved in the interaction [77].

According to Salim et al. [77], measurements provide detailed descriptions and contextual information:

- Measurements related to the interaction itself (pink box, e.g., `ppi:resourceCollected`, `ppi:nectarCollectingBodyPart`, `ppi:numberOfRemovedPollenGrains`) are connected to the event via `dwc:eventID`, using the `dwc:MeasurementOrFact` class. These describe ecological or behavioral traits of the interaction.
- Measurements related to the individual organisms (green boxes, e.g., `ppi:flowerColor`, `ppi:floralAttractants`, `ppi:caste`) are linked using both `dwc:eventID` and `dwc:occurrenceID` via the `obis:ExtendedMeasurementOrFact` extension, enabling the schema to capture one-to-many relationships and rich organism-level data.

The direction and type of interaction (e.g., pollination, nectar collection) are represented using the `dwc:ResourceRelationship` class (orange box). This includes terms such as `dwc:resourceID` (the actor in the interaction), `dwc:relatedResourceID` (the recipient), and `dwc:relationshipOfResourceID` (the nature of the interaction, ideally drawn from a controlled vocabulary like the Relations Ontology). These relationships are also linked to the interaction event via `dwc:eventID` [77].

The PPI Vocabulary integrates multiple controlled vocabularies. Its flexible structure enables interoperability while covering the complexity of ecological interactions. Recognizing its broader applicability, efforts are underway to align the PPI model with GBIF's Unified Data Model [271]. Both models share a common structure centered around the `dwc:Event` class, reinforcing a standardized approach to representing biotic interactions [78].

11.4.2 Controlled Vocabularies

Currently, DwC offers controlled vocabularies for only a few metadata fields, specifically `dwc:establishmentMeans`, `dwc:degreeOfEstablishment`,

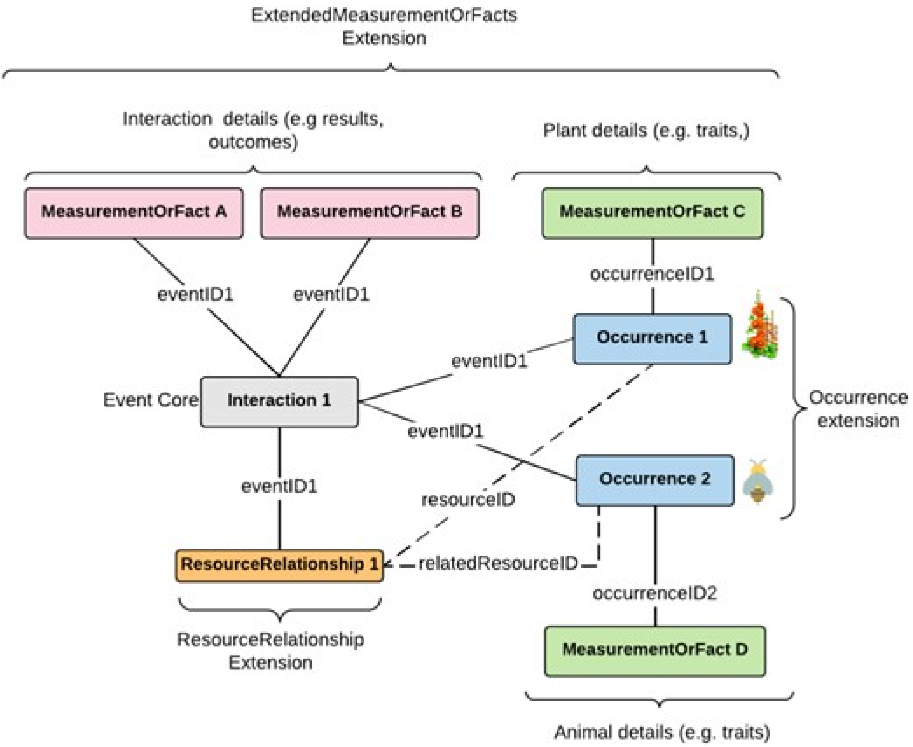


Figure 11.1 Overview of the Data Schema for Plant–Pollinator Interactions [77].

and `dwc:pathway` [272]. Although many other fields are intended to use controlled terms, the schema does not provide corresponding vocabularies. One notable exception is the term `dwc:relationshipOfResourceID`, for which DwC recommends using the OBO Relation Ontology (RO) [273]. This recommendation was introduced in a recent update to DwC following requests from projects like GloBI, Arctos, iNaturalist, and TrIAS, which had already been using RO to describe biotic relationships [78].

In this case study, RO was adopted by the project team to describe interaction types in a machine-readable and standardized way. Natural language terms were mapped to RO properties such as `ro:hasFlowersVisitedBy` and `ro:visitsFlowersOf`, allowing for clear and directional representation of plant-pollinator interactions [79].

Another important resource for the harmonization of agriculture data adopted in this case study is AGROVOC, which has also been used in Case Study 1. In the plant-pollinator interactions context, AGROVOC plays an important role in the characterization of crop plants' traits and interaction outcomes. In the PPI vocabulary, AGROVOC terms such as "ornamental plants" and "fuel crops" are used as controlled vocabulary for the PPI term `ppi:humanUse` [78].

11.4.3 Metadata Annotation

Metadata harmonization was essential to ensure the FAIRness of the datasets. To achieve this, metadata related to the datasets – such as publication date, reference citation, taxonomic range, and geographic location – was structured in a metadata table and converted into an EML metadata record. While EML is widely used, it is represented as an XML schema, which poses limitations for machine reasoning and semantic applications [80].

To solve this problem, we defined a mapping between EML and metadata schemas defined in RDF, such as Schema.org and DCAT. The metadata elements included in this mapping were based on the recommendations for cross-domain interoperability proposed by the CDIF Working Group [104]. This mapping, published on GitHub[1], allows datasets to be represented in alternative metadata schemas while maintaining compatibility with platforms like GloBI and GBIF, which require an EML metadata record for proper data integration [80].

The EML schema facilitated integration with GloBI, since the GloBI crawler recognizes EML as a valid format for metadata. By structuring metadata through EML and other schemas, this step ensured that plant-pollinator datasets could be searched, retrieved, and reused in a machine-readable format, enhancing their FAIRness and long-term usability [79].

11.4.4 Publishing the Annotated Datasets

The pilots were supported by various organizations, including private companies (e.g., HiveTracks), public institutions (e.g., USDA), and academic research centers (e.g., Universidad CES). Given the heterogeneity of the data sources and institutional policies, different data integration and access models were necessary. While some datasets were made fully open-access, others imposed access restrictions, particularly when linked to ongoing research (e.g., unpublished theses) [79].

The project team promoted discussions with the pilot projects on data-sharing strategies, highlighting the benefits and trade-offs of different data access models. FAIRified datasets with open access maximize visibility, community engagement, and external contributions, but sometimes restricted access to datasets is necessary due to institutional agreements and confidentiality requirements (e.g., datasets that have

---

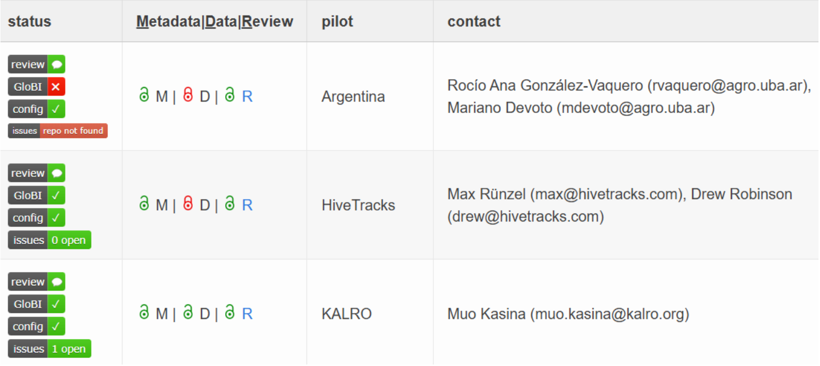
[1] `https://github.com/globalbioticinteractions/carvalheiro2023/issues/1#issuecomment-1855661190`

not yet been described in a publication). Recognizing this diversity, we emphasize it was necessary to understand institutional policies and historical contexts when integrating plant-pollinator data [79].

To improve the findability and accessibility of the datasets, a dedicated WorldFAIR webpage was launched on GloBI[2]. This webpage offers:

- A searchable index of species interaction data.
- Versioned pilot project metadata.
- Data review reports (Deliverable 10.3 [80]).
- A pilot-specific issue tracker for community discussions.

All datasets from the six pilots were classified by the accessibility of their metadata, data, and review reports, promoting transparency around what is openly available and what remains restricted. For example, as illustrated in Fig. 11.2, pilots from Argentina and HiveTracks provide open metadata and review reports, but restrict access to the actual data. In contrast, the KALRO pilot has made all components openly available. Contact information for each pilot is also provided, allowing users to request further details or negotiate access to restricted datasets. All annotated datasets were assigned DOIs and published on Zenodo. A complete list is available in Appendix III of [80] and in the Data and Code Availability section of this dissertation (datasets 20 to 29).



status	Metadata|Data|Review	pilot	contact
review 💬 GloBI ✗ config ✓ issues repo not found	🔓 M | 🔒 D | 🔓 R	Argentina	Rocío Ana González-Vaquero (rvaquero@agro.uba.ar), Mariano Devoto (mdevoto@agro.uba.ar)
review 💬 GloBI ✓ config ✓ issues 0 open	🔓 M | 🔒 D | 🔓 R	HiveTracks	Max Rünzel (max@hivetracks.com), Drew Robinson (drew@hivetracks.com)
review 💬 GloBI ✓ config ✓ issues 1 open	🔓 M | 🔓 D | 🔓 R	KALRO	Muo Kasina (muo.kasina@kalro.org)

Figure 11.2 Access Levels for WorldFAIR Pilots on the GloBI Webpage, Adapted from [79].

11.5 Validation

This section discusses the Validation process of the semantic interoperability framework (Section 2.6). In this case study, two types of

---

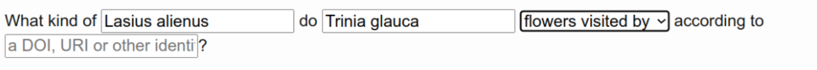
[2] `https://www.globalbioticinteractions.org/worldfair/`

validation protocols were implemented: a protocol that uses queries to retrieve data from one of the pilot projects, and one that leverages bots to crawl the data and generate a data review report.

11.5.1 Queries for FAIR Assessment

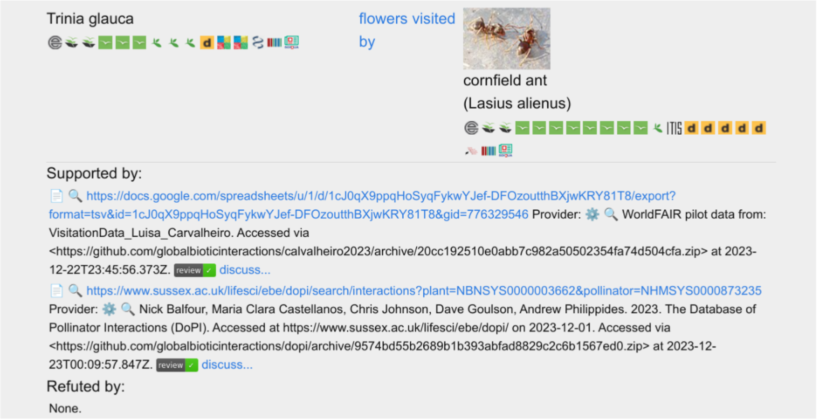
The purpose of querying the data was to confirm that its semantic integrity was preserved throughout the entire processing pipeline. Unlike traditional integration efforts (such as in Case Study 1, Section 9), this case study did not require data consolidation into a single platform. Instead, GloBI served as a federated search engine, accessing FAIR data access points on GitHub rather than storing datasets.

To illustrate this process, we have taken the Carvalheiro et al. [274] dataset as an example. Once the annotated dataset was uploaded to a GitHub repository tagged for GloBI indexing, automated review routines ensured correct mapping configurations and dataset accessibility. A query using GloBI[3] filters was performed, as shown in Fig. 11.3a, to confirm successful indexing and data retrieval, with results of this query shown in Fig. 11.3b.



What kind of Lasius alienus do Trinia glauca flowers visited by according to a DOI, URI or other identi ?

a Query for *Lasius alienus* visiting *Trinia glauca*.



Trinia glauca

flowers visited by

cornfield ant
(Lasius alienus)

Supported by:

https://docs.google.com/spreadsheets/u/1/d/1cJ0qX9ppqHoSyqFykwYJef-DFOzoutthBXjwKRY81T8/export?format=tsv&id=1cJ0qX9ppqHoSyqFykwYJef-DFOzoutthBXjwKRY81T8&gid=776329546 Provider: WorldFAIR pilot data from: VisitationData_Luisa_Carvalheiro. Accessed via <https://github.com/globalbioticinteractions/calvalheiro2023/archive/20cc192510e0abb7c982a50502354fa74d504cfa.zip> at 2023-12-22T23:45:56.373Z. review ✓ discuss...

https://www.sussex.ac.uk/lifesci/ebe/dopi/search/interactions?plant=NBNSYS0000003662&pollinator=NHMSYS0000873235 Provider: Nick Balfour, Maria Clara Castellanos, Chris Johnson, Dave Goulson, Andrew Philippides. 2023. The Database of Pollinator Interactions (DoPI). Accessed at https://www.sussex.ac.uk/lifesci/ebe/dopi/ on 2023-12-01. Accessed via <https://github.com/globalbioticinteractions/dopi/archive/9574bd55b2689b1b393abfad8829c2c6b1567ed0.zip> at 2023-12-23T00:09:57.847Z. review ✓ discuss...

Refuted by:

None.

b Search results from GloBI's index.

Figure 11.3 Query and Search Results for *Lasius alienus* visiting *Trinia glauca*, retrieved from GloBI's Search Index (2024-01-16) [79, p.43].

---

[3] https://www.globalbioticinteractions.org/

A key feature of GloBI, as shown in Fig. 11.3, is its ability to display all indexed datasets that support or refute a queried biotic interaction. This functionality allows users to assess both confirmations and contradictions in the data. In this example, no dataset disputed the interaction documented by Carvalheiro et al. [274]. Additionally, the Database of Pollinator Interactions (DoPI) [275] also referenced this interaction. By tracing DoPI's provenance, the original study by Carvalheiro et al. [276] was reused by that dataset, highlighting the importance of data reuse in biotic interaction research.

11.5.2 Automated FAIR-Enabled Data Reviews

Peer review ensures scientific quality of papers, but journal recommendations for reviewing datasets are often unclear or absent. Many journals do not provide explicit guidelines for data structure or quality, leading to the publication of unstructured, less reusable, and non-interoperable datasets [80].

This case study introduced an automated approach for reviewing plant-pollinator datasets, akin to a "spell-check" for scientific data. These automated reviews assist researchers at various stages of publication, flagging inconsistencies and enhancing data quality. By implementing machine-readable verification processes, this approach streamlines the review workflow while reinforcing the FAIR principles [80].

The automated data review follows the key steps below [80]:

1. Researchers create and register their plant-pollinator datasets according to GloBI guidelines[4] (outlined in Appendix IV of [80]).
2. A GloBI review bot automatically generates a human-readable report for evaluation (see Appendix V of [80]).
3. The researcher or peers review the report to assess potential issues.
4. If necessary, the dataset is revised and submitted for another review cycle.

This automated validation aligns with the FAIR principles by ensuring datasets are findable, accessible, interoperable, and reusable also for machines. The review process verifies elements such as the specificity of interaction types (e.g., pollination, flower visitation) and whether species names are consistently recognized across taxonomies like the Catalogue of Life[5] [80].

For this study, the Nomer Corpus of Taxonomic Resources [277] was used, incorporating versioned copies of multiple taxonomic databases,

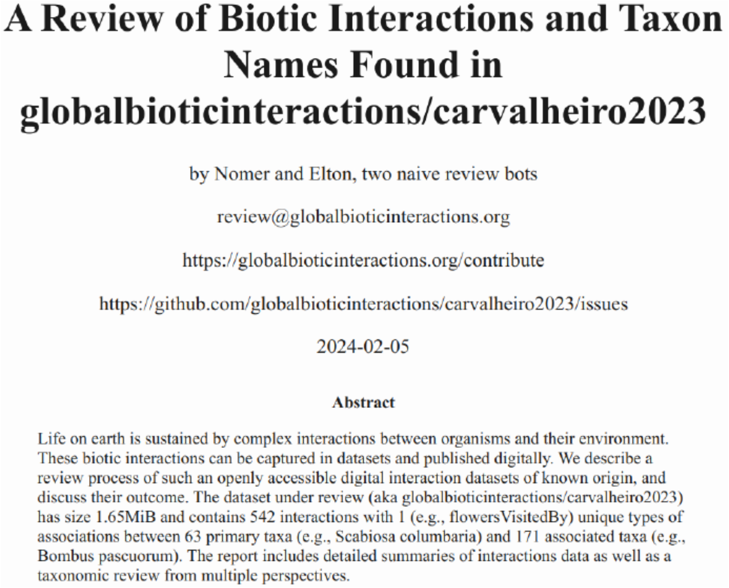
---

[4] See GloBI's data contribution guide at `https://www.globalbioticinteractions.org/contribute` and the "Guidelines and Recommendations for Publishing Agricultural-related Pollinator Data" [266].

[5] `https://www.catalogueoflife.org/`

including the GBIF Backbone Taxonomy, ITIS, NCBI Taxonomy, Index Fungorum, and World Flora Online [80].

For example, a review of Carvalheiro et al. [274] was generated on February 5, 2024, using Nomer and Elton, two GloBI review bots. To enhance usability, these bots were programmed to structure the report like a traditional scientific paper, complete with a title, authorship, abstract, and introduction, as shown in Fig. 11.4.



A Review of Biotic Interactions and Taxon Names Found in globalbioticinteractions/carvalheiro2023

by Nomer and Elton, two naive review bots

review@globalbioticinteractions.org

https://globalbioticinteractions.org/contribute

https://github.com/globalbioticinteractions/carvalheiro2023/issues

2024-02-05

Abstract

Life on earth is sustained by complex interactions between organisms and their environment. These biotic interactions can be captured in datasets and published digitally. We describe a review process of such an openly accessible digital interaction datasets of known origin, and discuss their outcome. The dataset under review (aka globalbioticinteractions/carvalheiro2023) has size 1.65MiB and contains 542 interactions with 1 (e.g., flowersVisitedBy) unique types of associations between 63 primary taxa (e.g., Scabiosa columbaria) and 171 associated taxa (e.g., Bombus pascuorum). The report includes detailed summaries of interactions data as well as a taxonomic review from multiple perspectives.

Figure 11.4 Example of the First Page of a Data Review Generated by GloBI Bots “Nomer” and “Elton” on 5 February 2024 [278] cited by [80].

Carvalheiro et al. [274] opted for open access to their data review, allowing the automated assessment results to be publicly shared. For datasets with restricted access, review workflows were adapted accordingly. A complete list of reviewed datasets and associated reports is available in Appendix II of Drucker et al. [80].

11.5.3 Taxonomic Name Alignment

Accurate taxonomic naming is essential for interpreting species interactions, but inconsistencies can arise due to typos (e.g., *Apis mellifera* vs. *Apis melifera*) or variations in species names (*Apis mellifera* vs. *Apis mellifera* Linnaeus). While automated tools can detect these discrepancies, they are often challenging for human reviewers to identify. For instance, in the review report for Kasina et al. [279], *Apis mellifera* appeared under four different names, highlighting the need for systematic name validation [78].

Comparing multiple taxonomic resources can reveal biases and inconsistencies in species alignments. The recognition of names varies across taxonomic catalogs, sometimes omitting valid names of rare or endemic species. For example, a plant dataset from Brazil benefits from a regional checklist, which enhances accuracy by identifying more valid names. This reveals the need for integrating both regional and international taxonomic catalogs to ensure interoperability across datasets [78].

Incorporating information from taxonomic revision papers can further refine species classification. A relevant case is the cyclanth pollinators, where one dataset correctly identifies a species (e.g., GBIF species 124558104), while another struggles with name mismatches at higher taxonomic levels (e.g., GBIF occurrence 1932562261) [78].

Although biologists understand the limitations of existing taxonomic catalogs, biodiversity informatics infrastructures do not always support their preferences. The ability to compare and align taxonomy from multiple perspectives is a crucial feature that needs further integration. Our automated review process demonstrates how this alignment can be effectively implemented, improving taxonomic accuracy across biodiversity datasets [78].

11.5.4 FAIR Rubrics for Plant-Pollinator Data

To enable the self-assessment of plant-pollinator interaction datasets, we developed a rubric for the evaluation process. This rubric consists of 11 key questions, each aligned with a FAIR Implementation Profile (FIP) [280] and corresponding FAIR principles. Users can access additional details through hyperlinks provided in Appendix I of [80].

Designed as a practical guide, the rubric suggests strategies for improving the FAIRness of plant-pollinator data. Each question includes possible responses to assist interpretation, recognizing that some datasets may require continuous improvement [80].

While the rubric offers a structured framework for evaluation, ensuring data quality remains the responsibility of researchers. They must establish clear data management practices to enhance findability, accessibility, interoperability, and reusability. Ultimately, applying FAIR principles effectively depends on collaboration and adherence to community standards [80].

The rubric has certain limitations. First, FAIR assessments are inherently qualitative and rely on self-reported compliance, which may introduce bias or inconsistencies across different datasets. Second, while the rubric aligns with existing FAIR principles and the FIPs, it does not account for domain-specific challenges that may require tailored solutions. For instance, some plant-pollinator datasets may lack metadata fields due to variations in data collection methodologies, making full compliance difficult. Third, the rubric focuses on data structure and accessibility but does not explicitly measure data accuracy, completeness, or ecological relevance, which are critical for meaningful reuse. Lastly, maintaining FAIRness over time is another challenge, as metadata and data hosting platforms may change, potentially affecting long-term accessibility and interoperability. Addressing these limitations requires ongoing community engagement, refinement of assessment criteria, and integration with automated validation tools to ensure that FAIR compliance remains practical and effective.

11.6 Comparative Analysis of Case Studies 1 and 2

To compare the two case studies, we analyze them through the lens of the framework parts introduced in Chapter 2, and summarized this comparison in Table 11.1. In Case Study 1, the goal was to achieve semantic interoperability among agricultural price index datasets from three Brazilian institutions – CEPEA, CONAB, and Ipea – with the objective of building an integrated KG. In contrast, Case Study 2 focused on harmonizing datasets from six international pilot projects to enable federated search via the GloBI platform. The data providers in Case Study 2 were more geographically and institutionally diverse, including the Facultad de Agronomía at Universidad de Buenos Aires (Argentina), REBIPP (Brazil), KALRO (Kenya), HiveTracks (USA), USDA (USA), Universidad CES, and SIB Colombia (Colombia). A key strength of Case Study 2 was the close collaboration with data providers: representatives from each institution actively participated in meetings, shared insights on their data management challenges, and helped refine the developed protocols. In contrast, in Case Study 1, only CEPEA was actively engaged. CONAB and Ipea did not attend meetings organized during Brazil's 5th National Action Plan on Open Government and provided minimal feedback on data clarification requests, as discussed in Section 4.4.1.

Table 11.1 Comparison of Case Studies 1 (Agricultural Price Index Data) and 2 (Plant-Pollinator Interactions) based on the Semantic Interoperability Framework.

Framework Step	**Case Study 1**	**Case Study 2**
Goal Definition	Integration of agricultural price index data from CEPEA, CONAB, and Ipea to build a national KG.	Harmonization of plant-pollinator interaction datasets from six international pilots to support federated search via GloBI.
Stakeholder Engagement	Limited; only CEPEA participated actively.	Strong engagement from all pilots; continuous feedback helped shape project outcomes.

Table 11.1 (continued)

Framework Step	**Case Study 1**	**Case Study 2**
Metadata Template	Custom list of terms, later formalized as the Almes Core schema.	Used the REBIPP template, built on EML, DwC, and PPI schemas.
Data Preparation	Required extensive cleaning (e.g., merged cells, logos in spreadsheets).	Datasets already in tabular format; minimal cleaning needed.
Data Completion	Metadata had to be requested via government platform (e.g., LAI for CONAB), with limited responses.	Pilots readily contributed missing metadata and participated in discussions.
Semantic Annotation	Developed new artifacts: Almes Core (metadata) and APTO (ontology).	Reused existing vocabularies: DwC, EML, PPI, RO, AGROVOC.
Semantic Web Compatibility	Almes Core and APTO designed in RDF/OWL, fully Semantic Web-ready.	EML, DwC, and PPI not designed for Semantic Web; represented as XML/tag-based structures.
Validation	Queries run on unified KG platform to validate integration.	Validation via GloBI federated search and bot-generated FAIR reviews.
FAIR Assessment	Used existing automated FAIR assessment tool.	Combined self-assessment rubric with machine-readable FAIR review reports (Nomer and Elton).
Taxonomic Data Use	Species linked to agricultural products; names reviewed via ChatGPT.	Species used to describe interactions; bots detected misspellings and inconsistencies.
AI Integration	ChatGPT supported ontology design and data querying.	GloBI bots (Nomer and Elton) used for review, validation, and name alignment.

For the data template, Case Study 1 utilized a list of essential terms, which later led to the development of Almes Core, since no specific metadata schema for this domain was available. In Case Study 2, the REBIPP template was proposed based on existing metadata schemas, including DwC, EML, and the PPI vocabulary.

The second process, Data Preparation, consisted of formatting the datasets according to the templates defined in the previous process. The first activity was to ensure that all deployed datasets were in the same format, namely tabular data.

In Case Study 1, this required several data manipulations, particularly with the CONAB's dataset. This dataset included elements such as institutional logos embedded in cells and numerous merged cells, which needed to be cleaned and reformatted. In contrast, all datasets in Case Study 2 were already in tabular format, simplifying the preparation process.

Filling out the templates required further data collection in both case studies. It was necessary to contact the data providers via email to request additional information. Specifically, for CONAB, the request had to be made through Brazil's Access to Information Law 12.527/2011 on a government platform. Despite following the proper protocol, the response was incomplete, which resulted in metadata records with no descriptions nor the start date of the data time series. As a federal institution, CONAB is obligated to share public-interest data unless it violates data protection rules. The data we requested was not sensitive; it was metadata needed to complete the templates. This experience highlights the challenges of obtaining complete data and metadata, even from public institutions, underscoring the importance of the open data and FAIR movements.

The most significant differences between the two case studies are found in the last two processes of the framework.

In the Semantic Annotation Process, Case Study 1 lacked both a domain-specific metadata schema and an ontology for annotating price index data. To cope with this we developed Almes Core and APTO. In contrast, Case Study 2 had both suitable domain metadata schemas (EML, DwC, PPI) for dataset annotation and an existing ontology (RO, AGROVOC) for encoding controlled vocabulary fields that describe biological interactions. This made Case Study 2 more straightforward.

However, there is an important limitation regarding the metadata schemas used in Case Study 2: neither EML, DwC, nor PPI were designed as Semantic Web schemas. EML is structured as a list of XML tags, and both DwC and PPI follow a similar structure. This lack of formal Semantic Web representation affects the reuse of the schemas, particularly in terms of interoperability, automated reasoning, and linking across different data sources. In contrast, Almes Core and APTO were designed as Semantic Web vocabularies, following the recommendations of the FAIR principles, as discussed in Chapter 3.

The second key difference between the case studies concerns the validation process. In Case Study 1, validation involved testing the integrated data from different providers by querying it through a unified platform. In contrast, Case Study 2 focused on federated search on GloBI, by checking whether FAIRified datasets made available via FAIR access points on GitHub, could be retrieved through GloBI's search engine.

The second part of Case Study 2's validation involved testing if two bots could crawl the data from these FAIR access points and generate a data review report. This process worked smoothly for all datasets, serving as a form of FAIR assessment. Only datasets that adhered to the FAIR data principles, as defined in the FAIR rubrics for plant-pollinator interaction data, were included in the project [80]. These rubrics were conceived as a manual approach to assessing the FAIRness of plant-pollinator datasets. In Case Study 1, the FAIR assessment was conducted using an existing automated tool, and we reviewed and commented on the results.

In both case studies, taxonomic data played a crucial role in identifying the species included in the datasets. In Case Study 1, taxonomic information was used to identify the species responsible for producing a given agricultural product. The relationship between product type and species within the ontology was key to resolving ambiguities arising from popular or regional names. In Case Study 2, species names were critical in indicating which species participate in specific interactions.

Both case studies faced a common challenge with taxonomic data coming from errors introduced through human manipulation, particularly typos. To address this issue, AI-based solutions were proposed in both cases. In Case Study 1, ChatGPT was employed to review the species list before processing it through the algorithm. In Case Study 2, GloBI's bots were trained to detect misspellings and typos in scientific names. These applications demonstrate the efficiency of AI in mitigating human errors during data manipulation.

In summary, this study underscores the importance of advancing semantic interoperability to enable seamless data integration and reuse. Both case studies demonstrate how challenges in this area can be addressed, either through the creation of new artifacts or by improving existing ones. The project highlights that true interoperability goes beyond data formatting – it requires a shared understanding of meaning across systems. Achieving this is key to unlocking the full potential of FAIR data practices, ensuring that diverse datasets can be efficiently used and accessed across platforms, thereby contributing to more robust, interconnected scientific research.

11.7 Final Considerations

This chapter outlined the application of a semantic interoperability framework to ensure the FAIRness of plant-pollinator interaction datasets across diverse pilot projects. By implementing standardized metadata schemas, ontologies, and automated review processes, the WorldFAIR project facilitated data integration and reuse while enhancing machine-actionability. The use of DwC, PPI, and EML enabled a structured approach to data annotation, while mapping metadata to broader schemas like Schema.org and DCAT ensured cross-domain compatibility.

Additionally, the GloBI indexing system played a key role in demonstrating how FAIR principles can be operationalized in real-world biodiversity informatics. Through querying, provenance tracking, and automated FAIR reviews, the framework supported transparent and accessible data integration. However, challenges remain, particularly regarding data access restrictions, taxonomic inconsistencies, and long-term sustainability of FAIR compliance.

The author of this dissertation was directly involved in the development of the FAIR self-assessment rubric and in the mapping of the metadata schemas adopted in the project (EML, DwC, and PPI) to cross-domain standards such as DCAT and Schema.org. These contributions supported the project's goal of enabling greater interoperability and reuse of plant-pollinator interaction datasets across institutional and disciplinary boundaries.

Moving forward, refining automated validation tools, strengthening taxonomy alignment, and fostering broader community engagement will be crucial for ensuring that plant-pollinator interaction data remains interoperable, reusable, and well-integrated within global biodiversity infrastructures.

CHAPTER 12

Final Remarks

This dissertation addressed the complex challenge of achieving semantic interoperability across heterogeneous agricultural datasets by proposing, developing, and validating a structured framework grounded in the FAIR data principles. Through the integration of conceptual modeling, metadata schema design, ontology development, and supporting tooling, this research offered practical solutions to a well-known yet persistently unresolved issue in data integration: the preservation of meaning across systems.

A key contribution was the articulation of a process-oriented semantic interoperability framework, which guided the step-by-step harmonization of data from disparate sources. This framework was tested and refined through two case studies. In the first, the goal was to semantically integrate agricultural price index datasets from Brazilian institutions; in the second, the focus was on enhancing the interoperability of plant-pollinator interaction data in collaboration with global partners through the WorldFAIR project. These case studies illustrated how technical strategies – such as the use of metadata schemas (e.g., Almes Core, EML, PPI, and DwC), ontologies (e.g., APTO, RO, AGROVOC), and automated review tools – can be effectively combined with stakeholder engagement and domain-specific knowledge to resolve structural and semantic mismatches.

Importantly, the execution of these case studies played a formative role in shaping the final structure of the framework. Several activities were added or revised in response to practical challenges encountered during implementation. For example, a previous version of the framework envisioned “Data Integration” as the sole validation activity. However, insights from Case Study 2 revealed the relevance of federated search in distributed data environments, prompting its inclusion as an additional validation strategy. This adaptive evolution underscores the framework’s flexibility and its grounding in real-world data interoperability needs.

Throughout the research, several recurring problems were identified and systematically addressed. These included terminological conflicts, data format inconsistencies, the absence of standardized metadata, lack of ontology reuse, and limited machine-actionability of datasets. These issues were resolved through the development of reusable semantic artifacts, the implementation of protocols for data cleaning and annotation, and the application of validation mechanisms such as SPARQL querying and FAIR self-assessment rubrics.

Ultimately, this work demonstrated that semantic interoperability is not a purely technical problem, but a sociotechnical one. It requires both formal modeling and community engagement, institutional coordination, and iterative refinement. The results of this research provide not only theoretical insights but also concrete tools and practices that can be applied to similar challenges in other domains.

12.1 Addressing the Semantic Interoperability Issues

In the introduction of this dissertation (Chapter 1), semantic interoperability was framed as the ability of information systems to exchange and interpret data in a way that preserves its meaning, structure, and context. A variety of known challenges were discussed, such as naming conflicts, inconsistencies in domain representation, mismatched levels of abstraction, lack of common vocabularies, and limited metadata reuse. These problems were theorized as barriers that often prevent effective data integration, reuse, and discovery –particularly in complex, heterogeneous domains such as agriculture and biodiversity.

As the research unfolded, some of these anticipated challenges were empirically confirmed in both case studies:

- **Naming Conflicts**: As observed in Case Study 1, different institutions (e.g., CEPEA and Ipea) used different terms (e.g., periodicidade vs. frequência) for the same concept. This is a classical naming conflict and was resolved by defining a unified metadata schema (Almes Core) where equivalent terms were mapped to a shared semantic reference.
- **Data Encoding Conflicts**: in Case Study 1, data encoding conflicts emerged when comparing date formats across datasets from Ipea, CEPEA, and CONAB. While CEPEA and CONAB adopted commonly used formats for dates (dd/mm/yyyy) Ipea used an inconsistent and often ambiguous format (yyyy.mm).
- **Domain Representation Conflicts**, such as abstraction mismatches and metamodel discrepancies, were resolved through ontological modeling. This was particularly important for ensuring consistency in how entities like product types or biotic interactions were

modeled across different contexts and granularity levels, such as the case of ‘Pimenta’ and ‘Pepper’ discussed in Section 7.6.

In summary, the dissertation not only confirmed the existence of the theoretical semantic interoperability challenges but also demonstrated how they manifest in real-world agricultural and biodiversity datasets. Each challenge was met with targeted strategies grounded in the proposed semantic interoperability framework, leading to tangible improvements in the findability, accessibility, interoperability, and reusability of the data.

12.2 Research Objectives and Outcomes

This dissertation aimed to explore a specific hypothesis: that implementing a framework integrating metadata schemas, conceptual models, and ontologies could significantly enhance semantic interoperability across agricultural datasets. By doing so, the research aimed to resolve conflicts in data descriptions and domain representation, enabling seamless data exchange and reuse.

In line with this, the main objective – to propose and implement a semantic interoperability framework aligned with the FAIR data principles – was both methodologically and practically fulfilled. The framework was formalized using BPMN and supported by a suite of novel artifacts, including metadata schemas (such as Almes Core), ontologies (like APTO), tools for semantic annotation, and KG construction workflows.

Each specific research goal contributed directly to testing and refining this framework:

- The development of a conceptual model and design guidelines for FAIR metadata schemas addressed a gap in the literature and proved valuable when designing Almes Core.
- The design and publication of semantic artifacts, including Almes Core and APTO, allowed for domain-specific modeling while reusing concepts from well-established vocabularies like DCAT, Schema.org, and AGROVOC/Agrotermos.
- The implementation of tools, such as the Taxonomy OWLizer and custom RDF conversion scripts, facilitated real-world semantic annotation tasks.
- The application of the framework to two distinct case studies – one focusing on agricultural price data in Brazil and another on international biodiversity data—demonstrated the framework’s flexibility and broader relevance.
- Finally, the evaluation of the framework through FAIR self-assessments, ontology usability testing, and queries confirmed

its practical utility, especially in enhancing the quality, reuse potential, and semantic clarity of agricultural data.

Overall, the results support the initial hypothesis and validate the design science approach adopted in this study. The dissertation also highlights the social and institutional challenges of semantic interoperability, confirming the relevance of treating the framework as a socio-technical artifact.

12.3 Contributions

This dissertation introduces key innovations in the domain of FAIR data management, semantic interoperability, and knowledge graph applications, with a particular focus on agriculture and biodiversity informatics. While the work is grounded in solving concrete, domain-specific problems, it also provides theoretical insights relevant to broader communities working with semantic technologies, data modeling, and socio-technical systems.

The main contributions of this research include:

1. Conceptual model for FAIR metadata schemas (Chapter 3): This is the first formal proposal of a structured model specifically tailored to guide the development of FAIR-compliant metadata schemas. It addresses a key gap in the FAIR data literature, which historically focused on datasets and repositories but lacked guidance at the schema level. This model supports both domain experts and data stewards in creating interoperable and reusable metadata schemas.
2. Almes Core (Section 5.2): The first metadata schema for agricultural price index datasets, Almes Core fills a crucial standardization gap in economic and agronomic data. By aligning with DCAT, Schema.org, and FAIR principles, this schema enables uniform data sharing across institutions and lays the foundation for integrated agricultural economic analysis..
3. APTO (Chapter 6), which formalizes the complex and heterogeneous classification of agricultural products in Brazil – one of the most biodiverse and agriculturally productive countries. APTO enables consistent semantic annotation of product data, supporting both national trade analysis and cross-border interoperability in agricultural knowledge bases.
4. Ontology engineering with LLM support (Chapter 7): An innovative approach that integrates LLMs into ontology design and refinement, demonstrating how AI-driven methodologies can improve the efficiency and accuracy of ontology development.
5. Taxonomy OWLizer web application (Chapter 7): A novel tool for transforming species names from the GBIF Backbone Taxonomy

into OWL ontologies, enabling researchers to generate taxonomic ontologies dynamically via API queries.

6. Adapted ontology usability evaluation protocol (Chapter 8): A tailored methodology for assessing the usability of ontologies, refining existing usability evaluation techniques to better suit domain-specific knowledge representation needs.
7. C4AI-KGAP: A semantic integration model that integrates agricultural price index data from three major Brazilian institutions, enabling more accurate economic analysis in agriculture.
8. Generative AI-assisted querying and validation of KGs (Chapter 10): An experimental approach that uses natural language prompts to generate SPARQL queries, streamlining data retrieval, validation, and analysis within semantic KGs.

In addition to these specific outputs, this dissertation advances theoretical understanding in three key areas:

1. **Semantic Interoperability as a Process-Oriented Challenge**: The research proposes and validates a process-oriented semantic interoperability framework, grounded in BPMN, that unifies technical and social activities. This model contributes to theory by conceptualizing interoperability as a dynamic, sociotechnical workflow that integrates metadata, ontologies, tooling, and stakeholder engagement.
2. **Design Science in Semantic Web Applications**: By combining a bottom-up approach for requirement elicitation with a top-down approach for foundational alignment, the research offers a hybrid methodological stance that supports both context-specific utility and theoretical abstraction. It illustrates how Design Science can be applied to balance empirical relevance and theoretical insight in the development of semantic artifacts.
3. **Socio-technical Knowledge Graph Design**: The work positions KGs not only as data integration tools but as socio-technical infrastructures that mediate meaning across institutions and languages. This reframing contributes to ongoing discourse about the epistemological role of KGs in scientific research and data governance.

In summary, this dissertation combines semantic web technologies, AI-assisted modeling, and structured metadata frameworks to address real-world problems while also offering theoretical contributions that extend beyond the immediate application context. It lays a foundation for further theorization and methodological development in the fields of semantic interoperability, ontology engineering, and FAIR data infrastructures.

12.4 Generalizability of Findings

While the semantic artifacts developed in this dissertation (e.g., APTO, Almes Core) are tailored to specific domains, we showed that the framework and methodology can be generalized. Designed to be domain-agnostic at its core, the framework integrates a modular pipeline, which includes conceptual modeling, metadata schema specification, ontology engineering, semantic annotation, and FAIR-aligned validation, which can be adapted to other fields facing similar interoperability challenges.

The generalizability of our framework and methodology is reinforced by its alignment with cross-domain standards such as DCAT, Schema.org, and the FAIR principles. The successful application of the same methodological pipeline across two distinct domains (agricultural economics and agrobiodiversity) offers empirical evidence of its adaptability. As interest in cross-domain data integration grows, particularly in areas like sustainability science, digital twins, and government data, the need for modular and FAIR-compliant semantic frameworks becomes increasingly pressing. This dissertation contributes such a reusable structure, designed for extension, validation, and reuse across a wide range of data ecosystems.

12.5 Limitations

While this dissertation presents a comprehensive framework for enhancing semantic interoperability in agricultural and biodiversity data, some limitations must be acknowledged, both in terms of scope and implementation. These limitations suggest areas where further refinement and future research are warranted.

Although the proposed framework was validated through two applied case studies, its evaluation was primarily conducted by the author and collaborators closely involved in the project. This self-applied evaluation represents a form of feasibility or utility validation, demonstrating how the framework can be effectively applied to real-world datasets. However, no formal empirical evaluation was conducted with independent third-party users. This poses a potential threat to the external validity of the findings, as the framework has not yet been tested by practitioners outside the research team. Future work should include such third-party assessments (ideally through interviews or usability testing with external domain experts) to further validate the framework.

For instance, while Case Study 2 benefited from the participation of international stakeholders who provided datasets and engaged in collaborative discussions, the actual data processing, annotation, and application of the framework were carried out by the WorldFAIR research team. As such, the influence of these partners, although valuable in shaping protocols and surfacing domain challenges, did not constitute an independent evaluation of the framework per se. Collecting formal feedback from these collaborators (e.g., through follow-up surveys or interviews) would strengthen the validation and could be considered in follow-up studies.

Moreover, while both case studies covered diverse datasets and demonstrated adaptability, broader testing in other agricultural subdomains (e.g., livestock, soil monitoring) or in entirely different disciplines (e.g., health, energy) would be required to generalize the framework's effectiveness more widely.

Semantic interoperability relies heavily on collaboration between data providers and ontology or metadata designers. In Case Study 1, limited engagement from some key data providers (i.e., Ipea and CONAB) posed challenges for clarifying data semantics and collecting missing metadata. This affected the completeness of the knowledge graph and demonstrated how institutional engagement remains a critical bottleneck in real-world semantic projects.

While several tools and scripts were developed, including the Taxonomy OWLizer app, SPARQL query generation using LLMs, and RDF converters, many activities within the framework still require significant human input. Ontology design, metadata mapping, and schema creation remain labor-intensive tasks that demand domain expertise. Full automation was not achieved and may not be desirable in all cases, but this limits scalability without additional investment in human resources.

Some components of the dissertation, such as the use of LLMs for ontology support and SPARQL query generation, are still in early exploratory phases. While promising results were obtained, LLMs occasionally produced incorrect outputs, misunderstood contextual information, or failed to adapt to specific named graph structures without explicit instruction. Further fine-tuning or domain-specific training would be required to reliably deploy these approaches in production environments.

Given that many of the datasets in Case Study 1 are in Portuguese and most semantic web standards (e.g., DCAT, Schema.org) use English as the default language, some limitations emerged around multilingual metadata alignment. Although some support was added through language tags and translations, this aspect remains a challenge for global reuse and cross-lingual interoperability.

Finally, the developed artifacts were published in open repositories. However, the long-term sustainability and governance of these resources remain open questions. Ensuring updates, versioning, and community stewardship will be essential for preserving their value and usability over time.

12.6 Future Work

While this dissertation advances the field of semantic interoperability in agriculture by proposing a practical framework, designing novel semantic artifacts, and applying them to real-world case studies, several avenues for future research remain open and promising.

One exciting direction is the integration of KGs with Retrieval-Augmented Generation (RAG) architectures, particularly Graph-RAG. Although this work explored the generation of SPARQL queries using LLMs, further research could investigate how structured knowledge in agricultural KGs can be leveraged in RAG pipelines to improve context-aware natural language generation. For instance, Graph-RAG could enhance question answering or decision support systems for agriculture by retrieving relevant graph substructures and using them to inform or constrain LLM outputs.

The current framework was validated with static datasets. Future work could extend this approach to handle streaming or dynamically updated agricultural data, such as live market prices or IoT sensor data from precision agriculture. This would involve defining FAIR workflows for real-time data processing, integrating versioning practices, and ensuring semantic consistency over time.

As ontologies such as APTO and metadata schemas like Almes Core are reused or extended, managing their evolution becomes critical. Future studies could propose automated or semi-automated tools for ontology versioning, change tracking, and alignment with external vocabularies. This includes mechanisms to notify data publishers and users about updates or deprecated terms, minimizing risks to interoperability. Future extensions of APTO and Almes Core should be strategically developed to support new projects and domains, ensuring broader applicability while maintaining semantic coherence with the original models.

Brazil's agricultural data landscape, as demonstrated in this work, includes terms in Portuguese that are often difficult to map directly to international vocabularies. Future research could explore semantic lifting – the process of enriching syntactic representations with formal semantics – as a means to improve this mapping. By lifting local terms into higher-level semantic concepts using multilingual ontologies, data can become more interoperable across linguistic and cultural boundaries. Additionally, integrating multilingual language models for term disambiguation and cross-language alignment could further support accurate mappings and enhance global data integration.

The thesis used both manual FAIR rubrics and machine-based review mechanisms for dataset evaluation. A promising direction involves developing FAIRness assessment systems that leverage reasoning over ontologies and metadata. Such tools could automatically infer missing elements, suggest improvements, or assess compliance with evolving standards, such as those proposed by the CDIF initiative.

Finally, future work should address the human side of semantic tools. Interfaces like the Taxonomy OWLizer app demonstrated that simple tools can enable semantic enrichment by non-specialists. Research on usability, explainability, and cognitive load in ontology-driven platforms could empower broader adoption by stakeholders such as farmers, agronomists, or policy makers.

Data and Code Availability

The datasets and codes used in this dissertation are archived and documented in the following repositories:

1. Soares, F. M., Corrêa, F. E., & Centro de Estudos Avançados em Economia Aplicada (CEPEA). (2024). Raw Data from Cepea on Sugar, Fed Cattle, Coffee, and Soybean Price Indexes [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.12163228`.
2. Soares, F. M., Corrêa, F. E., & Instituto de Pesquisa Econômica Aplicada (Ipea). (2024). Raw Price Index Data from IpeaData on Leather and Leather Goods, Cellulose Pulp, Paper and Paper Products, Tobacco Products, and Wood Products [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.12169699`.
3. Soares, F. M., Corrêa, F. E., & Companhia Nacional de Abastecimento (Conab). (2024). Raw Price Index Data from Conab on Sugar, Fed Cattle, Coffee, and Soybean [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.12170310`.
4. Soares, F. M. (2024). Excel Files Merger [Computer software]. Zenodo. `https://doi.org/10.5281/zenodo.12206148`.
5. Soares, F. M. (2024). Treated Price Index Data from Cepea, Ipea, and Conab [Dataset]. Version 2. Zenodo. `https://doi.org/10.5281/zenodo.12531081`.
6. Soares, F. M. (2024). transform-dates.py script [Computer software]. Zenodo. `https://doi.org/10.5281/zenodo.12532448`
7. Soares, F. M. (2024). Treated Price Index Data from Cepea, Ipea, and Conab [Dataset]. Version 3. Zenodo. `https://doi.org/10.5281/zenodo.12580972`
8. Soares, F. M., Corrêa, F. E., Moreira, D. de A., Pignatari Drucker, D., Braghetto, K. R., Botazzo Delbem, A. C., Ferreira Pires, L., Bonino da Silva Santos, L. O., & Saraiva, A. M. (2024). Agriculture and Livestock Metadata Elements Set (Almes Core) [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.12711290`

9. Soares, F. M, Ferreira Pires, L., Bonino da Silva Santos, L. O., Calhau, R. F., Coura Moreira dos Santos Maculan, B., Coyle, K., Wang, S., Folmer, E., Drucker, D., de Almeida Campos, M. L., Marcondes, C. H., Barcellos Almeida, M., Braghetto, K. R., Ataíde Dias, G., Salim, J. A., Correa, F. E., Abreu Moreira, D. de ., Botazzo Delbem, A. C., & Saraiva, A. M. (2025). Assessing the FAIRness of Metadata Schemas. 43rd International Conference on Conceptual Modeling (ER2024), Pittsburgh. Zenodo. `https://doi.org/10.5281/zenodo.14779549`
10. Soares, F. M., Saraiva, A. M., Ferreira Pires, L., da Silva Santos, L. O. B., de Abreu Moreira, D., Corrêa, F. E., Braghetto, K. R., Pignatari Drucker, D., & Botazzo Delbem, A. C. (2025). Supporting Data for “Exploring ChatGPT-4 for Transforming Taxonomic Data into OWL: Lessons Learned and Implications for Ontology Development” [Data set]. Version 2. Zenodo. `https://doi.org/10.5281/zenodo.14982527`
11. Soares, F. M. (2025). Taxonomy OWLizer. [Software]. Version 2. Zenodo. `https://doi.org/10.5281/zenodo.14984136`
12. Soares, F. M. (2024). GPT-4 full log of interactions. [Data set]. GitHub. Available at: `https://github.com/Filipi-Soares/Thesis/blob/GPT4interactions/GPTlog032024`
13. Soares, F. M., Ferreira Pires, L., Olavo Bonino da Silva Santos, L., Corrêa, F. E., de Abreu Moreira, D., Pignatari Drucker, D., Braghetto, K. R., Botazzo Delbem, A. C., & Mauro Saraiva, A. (2025). Agriculture Product Types Ontology (APTO) [OWL Code]. Version 1.31. Zenodo. `https://doi.org/10.5281/zenodo.15008549`
14. Soares, F. M. (2024). Usability Evaluation of the Agriculture Product Types Ontology (APTO) [Dataset]. Version 2. Zenodo. `https://doi.org/10.5281/zenodo.13932057`
15. Soares, F. M. (2024). The C4AI Knowledge Graph on Agricultural Prices (C4AI-KGAP) [Dataset]. Zenodo.`https://doi.org/10.5281/zenodo.13685709`
16. Soares, F. M. (2024). Metadata file in Turtle. GitHub. `https://github.com/Filipi-Soares/MetaID/blob/main/ID.ttl`
17. Soares, F. M. (2024). Scripts for converting CSV data to RDF/Turtle using Python. [Computer software]. Zenodo. `https://doi.org/10.5281/zenodo.13687647`
18. Soares, F. M. (2024). The C4AI Knowledge Graph on Agricultural Prices (C4AI-KGAP) (v1.2 - This version includes a file with SPARQL queries.) [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.13741165`

19. Carvalheiro, L. G., Soares, F. M., Salim, J. A., Poelen, J. H., & Drucker, D. (2024). Provenance of WorldFAIR pilot data from: VisitationData_Luisa_Carvalheiro. [Dataset]. Zenodo. `https://doi.org/10.5281/zenodo.10530109`
20. Alves, D. A. (2024). Contribution of insect pollinators to orange production and quality [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679727`
21. Bedoya Duque, C., Correa Álvarez, C. M., Cardona Duque, J., Molina, A., Juan Fernando, A., Vélez-Naranjo, M. C., Marentes Herrera, E., Agudelo Zapata, M. C., & Idárraga Giraldo, M. C. (2024). Web interactions between insects and some common plants in the ”Refugio de Vida Silvestre Alto de San Miguel” [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10678237`
22. Carvalheiro, L. G. (2024). Plant-flower visitor network from Avon Gorge, UK [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679322`
23. Ferreira Marques, B., & Carvalheiro, L. G. (2024). Orange (Citrus sinensis L. Osbeck, var. Pera-rio) insect floral visiting data of orchards in Itaberaí, Goiás, Brasil [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679666`
24. González-Vaquero, R. A., & Devoto, M. (2024). Plant-flower visitor interactions recorded in Argentina (Buenos Aires: Carlos Casares county) [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10669877`
25. Graham, C., Varassin, I., & Machado-de-Souza, T. (2024). The Ecology of Plant Hummingbird Interactions (EPHI) - Brazil [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679802`
26. Kasina, M., Kimani, I., Mulwa, S., Wafula Muliaro, J., & Kenyatta, J. (2024). A review of the status of web-based African Plant-Pollinator Interaction data [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679258`
27. Rünzel, M., & Robinson, D. (2024). HiveTracks WorldFAIR Test Data [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10692151`

28. Tinoco, C. (2024). Floral visitation in restored areas/remnants of natural vegetation in the Xingu region [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10679395`
29. Wolowski, M., Agostini, K., Freitas, L., Bergamo, P., & Salim, J. A. (2024). Data compiled from published (original or review) studies carried out in Brazil on the reproductive system and pollination/pollinators of crop plants. [Data set]. Zenodo. Edited by Drucker, Debora; Salim, José Augusto; Poelen, Jorrit; Soares, Filipi. `https://doi.org/10.5281/zenodo.10691993`

Author Publications

During his Ph.D., the author has contributed to top conferences in the domain, such as:

International Conference on Conceptual Modeling

The author was the co-chair of the workshop ”SmartFood 2023”, in Lisbon, Portugal:

Guizzardi, R., Faron, C., Miranda Soares, F. M., & Sedrakyan, G. (2023). Front Matter: SmartFood - 1st Workshop on Controlled Vocabularies and Data Platforms for Smart Food Systems. In *Advances in Conceptual Modeling: ER 2023 Workshops, CMLS, CMOMM4FAIR, EmpER JUSMOD, OntoCom, QUAMES, and SmartFood* Lisbon, Portugal, November 6–9, 2023 Proceedings (1 ed., pp. 295-297). (Lecture Notes in Computer Science; Vol. 14319). Springer. `https://doi.org/10.1007/978-3-031-47112-4`

Still in the ER edition, the author presented the paper:

Soares, F. M., Bergier, I., Coradini, M. C., Lüdtke Ferreira, A. P., Ambrosio Telles, M., Moreira dos Santos Maculan, B. C., Alencar, M. D. C. F., Marques Simão, V. P., Teixeira de Almeida, B., Pignatari Drucker, D., dos Santos Machado Vieira, M., & Serra da Cruz , S. M. (2023). Unveiling Knowledge Organization Systems’ Artifacts for Digital Agriculture with Lexical Network Analysis. In T. P. Sales, G. Guizzardi, J. Araújo, & J. Borbinha (Eds.), *Advances in Conceptual Modeling: ER 2023 Workshops, CMLS, CMOMM4FAIR, EmpER, JUSMOD, OntoCom, QUAMES, and SmartFood*, Lisbon, Portugal, November 6–9, 2023, Proceedings (1 ed., pp. 299-311). (Lecture Notes in Computer Science; Vol. 14319). Springer. `https://doi.org/10.1007/978-3-031-47112-4_28`

In the ER 2024, in Pittsburgh, PA, the author presented the paper:

Soares, F. M., Pires, L. F., Santos, L. O. B. D. S., Calhau, R. F., Moreira dos Santos Maculan, B. C., Coyle, K., Wang, S., Folmer, E., Pignatari Drucker, D., Campos, M. L. D. A., Marcondes, C. H., Almeida, M. B., Braghetto, K. R., Dias, G. A., Salim, J. A., Corrêa, F. E., de Abreu Moreira, D., Botazzo Delbem, A. C., & Saraiva, A. M. (2024). Towards a Conceptual Model for FAIR Metadata Schemas. In E. Gallinucci, & H. Yasar (Eds.), Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos : Co-located with ER 2024 (Vol. 3849, pp. 42-55). (CEUR Workshop Proceedings; Vol. 3849). CEUR. `https://ceur-ws.org/Vol-3849/forum4.pdf`

International Conference on Semantic Systems

In 2022, the author presented the following paper:

Soares, F. M., Corrêa, F. E., Pires, L. F., da Silva Santos, L. O. B., Drucker, D. P., Braghetto, K. R., de Abreu Moreira, D., Delbem, A. C. B., da Silva, R. F., da Silva Lopes, C. O., & Saraiva, A. M. (2022). Building a Community-Based FAIR Metadata Schema for Brazilian Agriculture and Livestock Trading Data. In SEMPDW 2022 (CEUR workshop proceedings; Vol. 3235). `https://ceur-ws.org/Vol-3235/paper26.pdf`

Biodiversity Information Standards (TDWG) Conference

The author made several contributions to this conference, from 2020 to 2025:

Soares, F. M., Saraiva, A. M., Ferreira Pires, L., Santos, L. O. B., Drucker, D. (2025). Taxonomy OWLizer: A Web Tool for Converting Taxonomic Data into OWL Ontologies. *Biodiversity Information Science and Standards*, 9. In press.

Soares, F. M., Pires, L. F., Bonino da Silva Santos, L. O., Drucker, D. P., Moreira, D. A., Corrêa, F. E., Braghetto, K. R., Delbem, A. C. B., Saraiva, A. M. (2025). ChatGPT as a Semantic Engineering Assistant: Lessons from Ontology Design in the Agricultural Biodiversity Domain. *Biodiversity Information Science and Standards*, 9. In press.

Drucker, D. P., Soares, F. M., Poelen, J., Salim, J. (2025). Machine-Actionable Metadata in Practice: Lessons from Automating FAIR Assessment in Plant-Pollinator Datasets. *Biodiversity Information Science and Standards*, 9. In press.

Drucker, D., Salim, J., Poelen, J., & Soares, F. M. (2024). Collectively Working towards Plant-Pollinator Interactions Data Interoperability and Reuse: Lessons Learned from the WorldFAIR Project. *Biodiversity Information Science and Standards*, 8, Article e141109. `https://doi.org/10.3897/biss.8.141109`

Soares, F. M., Koffler, S., Ghilardi-Lopes, N. P., Carvell, C., Chiazzese, J., Pignatari Drucker, D., Salim, J. A., Albertini, B. D. C., Nunes-Silva, P., & Saraiva, A. M. (2023). Flower-Insect Timed Counts (FIT Count): protocol adaptation and preliminary results in Brazil. *Biodiversity Information Science and Standards*, 7, Article e111545. `https://doi.org/10.3897/biss.7.111545`

Soares, F. M., Ferreira Pires, L., Garcia, M. C., Martins de Carvalho, A., Koffler, S., Ghilardi-Lopes, N. P., Rangel Silva, R., Moreira dos Santos Maculan, B. C., Bertolini, A. M., Rigote, G., Coradin, L., Montedo, U. B., Pignatari Drucker, D., Santiago, R., de Carvalho, M. C., da Silva Lima, A. C., Reis de Almeida, K., Mendonça de França, S. G., Elias Gabriel, H. D., ... Saraiva, A. M. (2023). Optimizing the Monitoring of Urban Fruit-Bearing Flora with Citizen Science: An Overview of the Pomar Urbano Initiative. *Biodiversity Information Science and Standards*, 7, Article e112009. `https://doi.org/10.3897/biss.7.112009`

Drucker, D. P., Salim, J. A., Trekels, M., Groom, Q., Parr, C., Soares, F. M., Agostini, K., Saraiva, A. M., Molloy, L., Hodson, S., & Gregory, A. (2022). Plant-pollinator Interaction Data: A case study of the WorldFAIR project. *Biodiversity Information Science and Standards*, 6, Article e94310. `https://doi.org/10.3897/biss.6.94310`

Salim, J. A., Zermoglio, P., Drucker, D., Soares, F., Saraiva, A., Agostini, K., Freitas, L., Wolowski, M., Rech, A. R., Maués, M. M., & Varassin, I. G. (2021). Plant-pollinator Vocabulary: a Contribution to Interaction Data Standardization. *Biodiversity Information Science and Standards*, 5. `https://doi.org/10.3897/biss.5.75636`

Soares, F., Saraiva, A., & Drucker, D. (2020). Linking Agrobiodiversity Data through Metadata Standards. *Biodiversity Information Science and Standards*, 4, Article e58928. `https://doi.org/10.3897/biss.4.58928`

Full List of Publications

The author has contributed to several high-impact journals, including IEEE Access, Data Intelligence, and GigaScience, among others. A complete list of the author's publications is available below.

1. Soares, F. M., Saraiva, A. M., Ferreira Pires, L., Drucker, D. P., Braghetto, K. R., Bonino da Silva Santos, L. O., de Abreu Moreira, D., Corrêa, F. E., Delbem, A. C. B. (2025). A Novel UX-Based Approach for Ontology Evaluation: Applying Tree Testing to the

Agricultural Product Types Ontology. *IEEE Access*, 13. https://doi.org/10.1109/ACCESS.2025.3595447

2. Soares, F. M., Saraiva, A. M., Ferreira Pires, L., Santos, L. O. B. D. S., de Abreu Moreira, D., Corrêa, F. E., Braghetto, K. R., Drucker, D. P., Delbem, A. C. B. (2025). Exploring a Large Language Model for Transforming Taxonomic Data into OWL: Lessons Learned and Implications for Ontology Development. *Data Intelligence*, 7(2). https://doi.org/10.3724/2096-7004.di.2025.0020
3. Soares, F. M., Saraiva, A. M., Ferreira Pires, L., Santos, L. O. B., Drucker, D. (2025). Taxonomy OWLizer: A Web Tool for Converting Taxonomic Data into OWL Ontologies. *Biodiversity Information Science and Standards*, 9. In press.
4. Soares, F. M., Pires, L. F., Bonino da Silva Santos, L. O., Drucker, D. P., Moreira, D. A., Corrêa, F. E., Braghetto, K. R., Delbem, A. C. B., Saraiva, A. M. (2025). ChatGPT as a Semantic Engineering Assistant: Lessons from Ontology Design in the Agricultural Biodiversity Domain. *Biodiversity Information Science and Standards*, 9. In press.
5. Drucker, D. P., Soares, F. M., Poelen, J., Salim, J. (2025). Machine-Actionable Metadata in Practice: Lessons from Automating FAIR Assessment in Plant-Pollinator Datasets. *Biodiversity Information Science and Standards*, 9. In press.
6. Soares, F. M., Pires, L. F., Santos, L. O. B. D. S., Calhau, R. F., Moreira dos Santos Maculan, B. C., Coyle, K., Wang, S., Folmer, E., Pignatari Drucker, D., Campos, M. L. D. A., Marcondes, C. H., Almeida, M. B., Braghetto, K. R., Dias, G. A., Salim, J. A., Corrêa, F. E., de Abreu Moreira, D., Botazzo Delbem, A. C., & Saraiva, A. M. (2024). Towards a Conceptual Model for FAIR Metadata Schemas. In E. Gallinucci, & H. Yasar (Eds.), Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos : Co-located with ER 2024 (Vol. 3849, pp. 42-55). (CEUR Workshop Proceedings; Vol. 3849). CEUR. https://ceur-ws.org/Vol-3849/forum4.pdf
7. Drucker, D., Salim, J., Poelen, J., & Soares, F. M. (2024). Collectively Working towards Plant-Pollinator Interactions Data Interoperability and Reuse: Lessons Learned from the WorldFAIR Project. *Biodiversity Information Science and Standards*, 8, Article e141109. https://doi.org/10.3897/biss.8.141109
8. Saraiva, A. M., Osório, F. S., Colaço, A. F., Pignatari Drucker, D., Mendiondo, E. M., Corrêa, F. E., Soares, F. M., Molin, J. P., Benso, M. R., Marques, P. A. A., da Silva, R. F., Miranda, S. H. G. D., Costa, W. F., & Delbem, A. C. B. (2024). A inteligência artificial na pesquisa agrícola. *Revista Universidade de São Paulo*, 2024(141). https://doi.org/10.11606/issn.2316-9036.i141p91-106

9. Soares, F. M., Pires, L. F., Garcia, M. C., Coradin, L., Ghilardi-Lopes, N. P., Silva, R. R., Martins de Carvalho, A., Gavai, A., Bouzembrak, Y., Moreira dos Santos Maculan, B. C., Koffler, S., Montedo, U. B., Pignatari Drucker, D., Santiago, R., Carvalho, M. C. P. D., da Silva Lima, A. C., Elias Gabriel, H. D., Mendonça de França, S. G., Reis de Almeida, K., ... Saraiva, A. M. (2024). Citizen science data on urban forageable plants: a case study in Brazil. *GigaByte*. `https://doi.org/10.46471/gigabyte.107`
10. Soares, F. M., Saraiva, A. M., Rangel Silva, R., Ferreira Pires, L., Ghilardi-Lopes, N. P., Montedo, U. B., Moreira dos Santos Maculan, B. C., Martins de Carvalho, A., Koffler, S., Souza, T., Coradin, L., Pignatari Drucker, D., Santiago, R., Garcia, M. C., Bouzembrak, Y., Gavai, A., & Lanza, T. (2024). Fruits of Collaboration: the contribution of the ‘Pomar Urbano’ initiative in promoting the monitoring, conservation, and use of fruit-bearing plants in Brazilian cities. Poster session presented at Planetary Health Summit & 6th Annual Meeting 2024, Petaling Jaya, Malaysia.
11. Soares, F. M., Ferreira Pires, L., Garcia, M. C., Bouzembrak, Y., Coradin, L., Ghilardi-Lopes, N. P., Rangel Silva, R., Martins de Carvalho, A., Moreira dos Santos Maculan, B. C., Koffler, S., Montedo, U. B., Pignatari Drucker, D., Santiago, R., Gavai, A., Peres de Carvalho, M. C., da Silva Lima, A. C., Elias Gabriel, H. D., Mendonça de França, S. G., Reis de Almeida, K., ... Saraiva, A. M. (2024). Leveraging citizen science for monitoring urban forageable plants. *GigaScience*, 13, 1-3. Article giae007. `https://doi.org/10.1093/gigascience/giae007`
12. Drucker, D. P., Salim, J. A., Poelen, J., Soares, F. M., Gonzalez-Vaquero, R. A., Ollerton, J., Devoto, M., Rünzel, M., Robinson, D., Kasina, M., Taliga, C., Parr, C., Cox-Foster, D., Hill, E., Maues, M. M., Saraiva, A. M., Agostini, K., Carvalheiro, L. G., Bergamo, P., ... Trekels, M. (2024). *WorldFAIR (D10.2) Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations*. `https://doi.org/10.5281/zenodo.10666593`
13. Drucker, D. P., Salim, J. A., Poelen, J., Soares, F. M., Gonzalez-Vaquero, R. A., Devoto, M., Ollerton, J., Kasina, M., Carvalheiro, L. G., Bergamo, P. J., Alves, D. A., Varassin, I., Tinoco, C. F., Rünzel, M., Robinson, D., Cardona-Duque, J., Idárraga, M., Agudelo-Zapata, M. C., Herrera, E. M., ... Saraiva, A. (2024). *WorldFAIR (D10.3) Agricultural biodiversity FAIR data assessment rubrics*. `https://doi.org/10.5281/zenodo.10719265`
14. Silva, R. R., & Soares, F. M. (2024). O design gráfico como ferramenta para a sustentabilidade: uma experiência do projeto pomar urbano. In N. Kuwahara, S. C. Mota, A. C. Almeida, G. P. Ribeiro, S. Bouzon, & I. W. A. Ferreira (Eds.), Anais do XV Congresso Brasileiro de

Pesquisa e Desenvolvimento em Design - P&D Design EDUA. `https://doi.org/10.29327/5457226.1-226`

15. Correa, F. E., Pignatari Drucker, D., Soares, F. M., Braghetto, K. R., Botazzo Delbem, A. C., Osório, F. S., & Saraiva, A. M. (2024). Redes e parcerias para abertura de dados agrícolas. In C. I. C. Mendes, J. D. S. D. A. Maranhão, & A. M. Saraiva (Eds.), Agricultura digital, agrodados e regulação (1 ed., pp. 65-76). Embrapa. `http://www.alice.cnptia.embrapa.br/alice/handle/doc/1170049`
16. Drucker, D. P., Cruz, S., Telles, M., Ferreira, A., Corrêa, F., Bertin, P., Marassi, L., Aquino, K., Bezerra, G., Cruz, P., & Soares, F. M. (2023). Desdobramentos da Implementação da Rede GO FAIR Agro Brasil no Biênio 2021-2023. In *Proceedings of the 14th Brazilian Congress of Agroinformatics*. `https://doi.org/10.5753/sbiagro.2023.26570`
17. Soares, F. M., Koffler, S., Ghilardi-Lopes, N. P., Carvell, C., Chiazzese, J., Pignatari Drucker, D., Salim, J. A., Albertini, B. D. C., Nunes-Silva, P., & Saraiva, A. M. (2023). Flower-Insect Timed Counts (FIT Count): protocol adaptation and preliminary results in Brazil. *Biodiversity Information Science and Standards*, 7, Article e111545. `https://doi.org/10.3897/biss.7.111545`
18. Mendes, C., Correa, F., Drucker, D., Soares, F. M., Saraiva, A., Delbem, A., Osório, F., Oliveira, C., & Maranhão, J. (2023). Governança de dados na agricultura digital: relato da experiência do C4AI. In *Proceedings of the 14th Brazilian Congress of Agroinformatics*. `https://doi.org/10.5753/sbiagro.2023.26559`
19. Soares, F. M., Koffler, S., Nunes-Silva, P., & Ghilardi-Lopes, N. P. (2023). O protocolo FIT Count como ferramenta de ensino e aprendizagem da polinização. In *Anais do Simpósio Brasileiro de Polinização*. `https://doi.org/10.29327/1286549.1-1`
20. Soares, F. M., Ferreira Pires, L., Garcia, M. C., Martins de Carvalho, A., Koffler, S., Ghilardi-Lopes, N. P., Rangel Silva, R., Moreira dos Santos Maculan, B. C., Bertolini, A. M., Rigote, G., Coradin, L., Montedo, U. B., Pignatari Drucker, D., Santiago, R., de Carvalho, M. C., da Silva Lima, A. C., Reis de Almeida, K., Mendonça de França, S. G., Elias Gabriel, H. D., ... Saraiva, A. M. (2023). Optimizing the Monitoring of Urban Fruit-Bearing Flora with Citizen Science: An Overview of the Pomar Urbano Initiative. *Biodiversity Information Science and Standards*, 7, Article e112009. `https://doi.org/10.3897/biss.7.112009`
21. Domene, S. M. Á., Agostini, K., Almeida, G. N. P. D., Camargo, R. G. M., Martins de Carvalho, A., Corrêa, F. E., Delbem, A. C. B., Domingos, S. S., Pignatari Drucker, D., Marchioni, D. M. L., Martins, I. P., Montedo, U. B., Ribeiro, E. M. S., Santiago, R., da Silva, R. F.,

Soares, F. M., Steluti, J., & Saraiva, A. M. (2023). Segurança alimentar: reflexões sobre um problema complexo. *Estudos Avancados*, 37(109), 181-206. `https://doi.org/10.1590/s0103-4014.2023.37109.012`

22. Soares, F. M., Bergier, I., Coradini, M. C., Lüdtke Ferreira, A. P., Ambrosio Telles, M., Moreira dos Santos Maculan, B. C., Alencar, M. D. C. F., Marques Simão, V. P., Teixeira de Almeida, B., Pignatari Drucker, D., dos Santos Machado Vieira, M., & Serra da Cruz , S. M. (2023). Unveiling Knowledge Organization Systems' Artifacts for Digital Agriculture with Lexical Network Analysis. In T. P. Sales, G. Guizzardi, J. Araújo, & J. Borbinha (Eds.), *Advances in Conceptual Modeling: ER 2023 Workshops, CMLS, CMOMM4FAIR, EmpER, JUSMOD, OntoCom, QUAMES, and SmartFood*, Lisbon, Portugal, November 6–9, 2023, Proceedings (1 ed., pp. 299-311). (Lecture Notes in Computer Science; Vol. 14319). Springer. `https://doi.org/10.1007/978-3-031-47112-4_28`
23. Trekels, M., Pignatari Drucker, D., Salim, J. A., Ollerton, J., Poelen, J., Soares, F. M., Rünzel, M., Kasina, M., Groom, Q. J., & Devoto, M. (2023). *WorldFAIR Project (D10.1) Agriculture-related pollinator data standards use cases report*. WorldFAIR Project. `https://doi.org/10.5281/zenodo.8176978`
24. Koffler, S., Acosta, A. L., Soares, F. M., & Saraiva, A. M. (2022). 2021 Planetary Health Annual Meeting and Festival Book of Abstracts: Planetary Health for All: Bridging Communities to Achieve the Great Transition. Universidade de Sao Paulo. `https://doi.org/10.5281/zenodo.6373367`
25. Soares, F. M., Hamanaka, R. Y., Fernandes Pontes, T. C., Araújo, W. J., & Moreira dos Santos Maculan, B. C. (2022). A aplicação do método de análise de conteúdo na ciência da informação: um estudo preliminar no contexto das teses e dissertações da UFMG. *Revista Ibero-Americana de Ciência da Informação*, 15(2). `https://doi.org/10.26512/rici.v15.n2.2022.36060`
26. Soares, F. M., Corrêa, F. E., Pires, L. F., da Silva Santos, L. O. B., Drucker, D. P., Braghetto, K. R., de Abreu Moreira, D., Delbem, A. C. B., da Silva, R. F., da Silva Lopes, C. O., & Saraiva, A. M. (2022). Building a Community-Based FAIR Metadata Schema for Brazilian Agriculture and Livestock Trading Data. In SEMPDW 2022 (CEUR workshop proceedings; Vol. 3235). `https://ceur-ws.org/Vol-3235/paper26.pdf`
27. Ghilardi-Lopes, N., Zattara, E., Agostini, K., Francoy, T., Fontúrbel, F., Soares, F. M., Blochtein, B., Koffler, S., Barbiéri, C., & Saraiva, A. M. (2022). Citizen science and pollinators of South America. In *Proceedings of the ECSA CONFERENCE 2022: Citizen science for planetary health* (pp. 34-34). European Citizen Science Association.

28. Salim, J. A., Saraiva, A. M., Zermoglio, P. F., Agostini, K., Wolowski, M., Drucker, D. P., Soares, F. M., Bergamo, P. J., Varassin, I. G., Freitas, L., Maués, M. M., Rech, A. R., Veiga, A. K., Acosta, A. L., Araujo, A. C., Nogueira, A., Blochtein, B., Freitas, B. M., Albertini, B. C., ... Brito, V. L. G. (2022). Data standardization of plant-pollinator interactions. *GigaScience*, 11, Article giac043. `https://doi.org/10.1093/gigascience/giac043`

29. Koffler, S., Soares, F. M., Ghilardi-Lopes, N. P., Albertini, B., Drucker, D., Salim, J. A., Nunes-Silva, P., Francoy, T. M., Saraiva, A. M., & Carvell, C. (2022). FIT Count Brasil: Aplicativo para monitoramento de polinizadores. In *II Workshop da Rede Brasileira de Ciência Cidadã* Even3. `https://doi.org/10.29327/175207.1-1`

30. Koffler, S., Soares, F. M., Ghilardi-Lopes, N. P., Albertini, B., Drucker, D. P., Salim, J. A., Nunes-Silva, P., Francoy, T. M., Saraiva, A. M., & Carvell, C. (2022). *FIT Count Brasil: monitoramento de visitantes florais por contagem.* (Série Ciência Cidadã; No. 7). UFABC. `https://doi.org/10.5281/ZENODO.6419201`

31. Carvell, C., Chiazzese, J., Zattara, E. E., Fontúrbel, F. E., Rivera, G. M., Ghilardi-Lopes, N. P., & Soares, F. M. (2022). Monitoramento da visitação de flores com Contagem Cronometrada de Visitantes Florais (FIT Count). In N. Pirani Ghilardi-Lopes, & E. E. Zattara (Eds.), *Ciência cidadã e polinizadores da América do Sul* (pp. 115-119). Editora Cubo. `https://doi.org/10.4322/978-65-86819-20-5.s03c14.pt`

32. Carvell, C., Chiazzese, J., Zattara, E. E., Fontúrbel, F. E., Rivera, G. M., Ghilardi-Lopes, N. P., & Soares, F. M. (2022). Monitoreo de visitas florales mediante el Conteo Cronometrado de Visitantes Florales (FIT Count). In N. P. Ghilardi-Lopes, & E. E. Zattara (Eds.), Ciencia ciudadana y polinizadores de América del Sur (1 ed., pp. 115-119). Editora Cubo. `https://doi.org/10.4322/978-65-86819-21-2.s03c14.es`

33. Soares, F. M., Koffler, S., Ghilardi-Lopes, N. P., Albertini, B., Nunes-Silva, P., Drucker, D. P., Carvell, C., Chiazzese, J., Salim, J. A., Francoy, T. M., & Saraiva, A. M. (2022). Monitoring pollinators in Brazil: the challenge of adapting the FIT Count citizen science protocol to the Brazilian context. In *Planetary Health Annual Meeting (PHAM 2022): Research, Project, & Lightning Talk Abstracts*, Planetary Health Alliance.

34. Saraiva, A. M., Leite, B. C. C., Albertini, B. D. C., Ulsen, C., Vianna, D., Miranda, D., Xavier, F., Soares, F. M., Borges, G. R., Machado, G., Grimoni, J. A. B., Jr, J. O., Ermilivitch, L., Vega, M. LBP., Okuyama, M. V. B., Seckler, M. M., Mauro, M. R., Pereira, M. C. S., Junior, M. A. S., ... Camargo, T. S. D. (2022). Planetary Health in Engineering:

building a community. In *Planetary Health Annual Meeting: Research, Project, & Lightning Talk Abstracts 2022*, Planetary Health Alliance.

35. Drucker, D. P., Salim, J. A., Trekels, M., Groom, Q., Parr, C., Soares, F. M., Agostini, K., Saraiva, A. M., Molloy, L., Hodson, S., & Gregory, A. (2022). Plant-pollinator Interaction Data: A case study of the WorldFAIR project. *Biodiversity Information Science and Standards*, 6, Article e94310. `https://doi.org/10.3897/biss.6.94310`
36. Soares, F. M., & Hamanaka, R. Y. (2021). Aplicação de metadados na padronização de registros de ocorrência de espécies no contexto da ciência cidadã para a biodiversidade: um estudo de caso. In *Organização do Conhecimento no Horizonte 2030: Desenvolvimento Sustentável e Saúde: Atas do V Congresso ISKO Espanha-Portugal*, Edicoes Colibri. `https://doi.org/10.51427/10451/50067`
37. Soares, F. M., Hamanaka, R. Y., & Moreira dos Santos Maculan, B. C. (2021). Interoperabilidade semântica no contexto de dados da biodiversidade: um estudo de caso sobre a utilização de padrões de metadados. In *XXI ENANCIB: Encontro Nacional de Pesquisa em Ciência da Informação*, Rio de Janeiro, 25 a 29 de outubro de 2021 Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação. `https://enancib.ancib.org/index.php/enancib/xxienancib/paper/view/79`
38. Salim, J. A., Zermoglio, P., Drucker, D., Soares, F., Saraiva, A., Agostini, K., Freitas, L., Wolowski, M., Rech, A. R., Maués, M. M., & Varassin, I. G. (2021). Plant-pollinator Vocabulary: a Contribution to Interaction Data Standardization. *Biodiversity Information Science and Standards*, 5. `https://doi.org/10.3897/biss.5.75636`
39. Soares, F., Saraiva, A., & Drucker, D. (2020). Linking Agrobiodiversity Data through Metadata Standards. *Biodiversity Information Science and Standards*, 4, Article e58928. `https://doi.org/10.3897/biss.4.58928`
40. Soares, F. M., Moreira dos Santos Maculan, B. C., Drucker, D. P., & Saraiva, A. M. (2020). Methodological principles to create a metadata extension to the Darwin Core standard for agrobiodiversity data. *Brazilian Journal of Information Science*, 14(4), e020015. `https://doi.org/10.36311/1940-1640.2020.v14n4.10865`
41. Soares, F. M., Moreira dos Santos Maculan, B. C., & Pignatari Drucker, D. (2020). Padrão de metadados Darwin Core: Proposta de extensão para as interações ecológi-cas no contexto da agrobiodiversidade. *Múltiplos Olhares em Ciência da Informação*, 9(2), 1-10. `https://periodicos.ufmg.br/index.php/moci/article/view/19179/`

42. Moreira, C., Soares, F. M., Hamanaka, R. Y., Bueno, R. V., & Aganette, E. C. (2020). Produtos da Ciência da Informação para o processo de doação de bens permanentes em Instituições Públicas: o caso do projeto Motirõ. *Múltiplos Olhares em Ciência da Informação*, 10. `https://doi.org/10.35699/2237-6658.2020.20344`

Bibliography

[1] Christof Ebert and Carlos Henrique C. Duarte. 'Digital Transformation'. In: *IEEE Software* 35.4 (2018), pp. 16–21. DOI: `10.1109/MS.2018.2801537`.

[2] Food and Agriculture Organization of the United Nations. *Digital Agriculture*. Accessed: July 31, 2023. URL: `https://www.fao.org/digital-agriculture/en`.

[3] A Shamin et al. 'Digital transformation of agricultural industry'. In: *IOP Conference Series: Earth and Environmental Science* 346.1 (Oct. 2019), p. 012029. DOI: `10.1088/1755-1315/346/1/012029`.

[4] S Santos Valle and Josef Kienzle. 'Agriculture 4.0: Agricultural robotics and automated equipment for sustainable crop production'. In: *Integrated Crop Management* 24 (2020). URL: `http://www.fao.org/3/cb2186en/CB2186EN.pdf`.

[5] Food and Agriculture Organization of the United Nations. *E-agriculture in action: Big data for agriculture*. Food and Agriculture Organization of the United Nations, International Telecommunication Union, 2019. URL: `https://openknowledge.fao.org/handle/20.500.14283/ca5427en`.

[6] A. M. Saraiva et al. 'Dados na Agricultura Digital: Ciclo, padronização, qualidade, compartilhamento e segurança'. In: *Agricultura Digital*. Viçosa, MG, Brazil: Universidade Federal de Viçosa (UFV), 2020, pp. 308–325.

[7] Anwaar Ali et al. 'Big data for development: applications and techniques'. In: *Big Data Analytics* 1.1 (Dec. 2016), p. 2. DOI: `10.1186/s41044-016-0002-4`.

[8] Mark D. Wilkinson et al. 'The FAIR Guiding Principles for scientific data management and stewardship'. In: *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: `10.1038/sdata.2016.18`.

[9] GO Fair. *Fair principles*. Accessed: August 1, 2023. URL: `https://www.go-fair.org/fair-principles/`.

[10] Mark D. Wilkinson et al. 'A design framework and exemplar metrics for FAIRness'. In: *Scientific Data* 5.1 (June 2018), p. 180118. DOI: `10.1038/sdata.2018.118`.

[11] Shirui Pan et al. 'Unifying Large Language Models and Knowledge Graphs: A Roadmap'. In: *IEEE Transactions on Knowledge and Data Engineering* (2024), pp. 1–20. DOI: `10.1109/TKDE.2024.3352100`.

[12] CrowdFlower. *2016 Data Science Report*. Tech. rep. CrowdFlower, 2016. URL: `https://www2.cs.uh.edu/~ceick/UDM/CFDS16.pdf`.

[13] IBM. *A glimpse inside the mind of a data scientist*. Tech. rep. IBM Analytics, 2016. URL: `https://www.ibm.com/downloads/cas/W6GEX9LL`.

[14] Steve Lohr. 'For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights'. In: *The New York Times* (Aug. 2014). URL: `https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html`.

[15] Subchefia para Assuntos Jurídicos Presidência da República Casa Civil. *Lei 12.527, de 18 de novembro de 2011*. 2011. URL: `https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`.

[16] Controladoria-Geral da União (CGU). *Modelo de Referência para Publicação de Dados Abertos*. CGU, 2020. URL: `https://repositorio.cgu.gov.br/handle/1/46701`.

[17] Controladoria-Geral da União (CGU). *4º Plano de Ação Nacional em Governo Aberto*. CGU, 2018. URL: `https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/4o-plano-de-acao-brasileiro/4o-plano-de-acao-nacional_portugues.pdf`.

[18] Office of the Comptroller General, Brazil. *Fifth National Action Plan on Open Government*. Brasília: Office of the Comptroller General, 2021. URL: `https://www.gov.br/cgu/pt-br/governo-aberto/a-ogp/planos-de-acao/5o-plano-de-acao-brasileiro/ing-5-plano-30-03-2022.pdf`.

[19] Debora Pignatari Drucker et al. 'Rumo à rede de implantação GO FAIR 'Agro' Brasil: a experiência de uma organização de PDI na implantação dos princípios FAIR'. In: *Princípios FAIR aplicados à gestão de dados de pesquisa*. Ibict, Oct. 2021, pp. 47–62. DOI: `10.22477/9786589167242.cap4`.

[20] Debora P. Drucker et al. ‘Implantação da Rede Temática GO-FAIR Agro Brasil: Primeiros Passos’. In: *Anais do XIII Congresso Brasileiro de Agroinformática (SBIAGRO 2021)*. Brasil: Sociedade Brasileira de Computação, Nov. 2021, pp. 164–171. DOI: `10.5753/sbiagro.2021.18387`.

[21] Sandra Heiler. ‘Semantic interoperability’. In: *ACM Computing Surveys* 27.2 (June 1995), pp. 271–273. DOI: `10.1145/210376.210392`.

[22] Peter Wegner. ‘Interoperability’. In: *ACM Computing Surveys* 28.1 (Mar. 1996), pp. 285–287. DOI: `10.1145/234313.234424`.

[23] A. M. Ouksel and A. Sheth. ‘Semantic interoperability in global information systems’. In: *ACM SIGMOD Record* 28.1 (Mar. 1999), pp. 5–12. DOI: `10.1145/309844.309849`.

[24] Tim Benson and Grahame Grieve. ‘Why Interoperability Is Hard’. In: *Principles of Health Interoperability: FHIR, HL7 and SNOMED CT*. Cham: Springer International Publishing, 2021, pp. 21–40. DOI: `10.1007/978-3-030-56883-2_2`.

[25] Moritz Lehne et al. ‘Why digital medicine depends on interoperability’. In: *NPJ Digital Medicine* 2 (Aug. 2019), p. 79. DOI: `10.1038/s41746-019-0158-1`.

[26] Marcia Lei Zeng. ‘Interoperability’. In: *KNOWLEDGE ORGANIZATION* 46.2 (2019), pp. 122–146.

[27] National Interoperability Framework Observatory (NIFO). *Interoperability layers*. URL: `https://joinup.ec.europa.eu/collection/nifo-national-interoperability-framework-observatory/3-interoperability-layers`.

[28] Federico Morando. ‘Legal interoperability: making Open Government Data compatible with businesses and communities’. In: *JLIS: Italian Journal of Library, Archives and Information Science* 4.1 (2013), pp. 441–452. DOI: `10.4403/jlis.it-5461`.

[29] International Electrotechnical Commission. *Semantic interoperability: challenges in the digital transformation age*. Oct. 2019. URL: `https://www.iec.ch/basecamp/semantic-interoperability-challenges-digital-transformation-age`.

[30] Bernhard Haslhofer and Wolfgang Klas. ‘A survey of techniques for achieving metadata interoperability’. In: *ACM Computing Surveys* 42.2 (Feb. 2010), pp. 1–37. DOI: `10.1145/1667062.1667064`.

[31] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 11179-1:2015 - Information technology - Metadata registries (MDR) - Part 1: Framework*. Tech. rep. ISO/IEC 11179-1:2015. Geneva, Switzerland: ISO/IEC, 2015.

[32] UNSD. 'Data and metadata models'. In: *Data Interoperability Guide*. United Nations Statistics Division, 2018. Chap. 2. URL: `https://unstats.un.org/wiki/display/InteropGuide/Chapter+2%3A+Data+and+metadata+models`.

[33] Giancarlo Guizzardi. 'Ontological Foundations for Structural Conceptual Models'. PhD thesis. University of Twente, 2005. URL: `https://research.utwente.nl/en/publications/ontological-foundations-for-structural-conceptual-models`.

[34] Giancarlo Guizzardi and Gerd Wagner. 'Conceptual simulation modeling with Onto-UML advanced tutorial'. In: *Proceedings of the 2012 Winter Simulation Conference (WSC)*. 2012, pp. 1–15. DOI: `10.1109/WSC.2012.6465133`.

[35] Natalia Fonseca Padilha, Fernanda Baião and Kate Revoredo. 'Ontology Alignment for Semantic Data Integration through Foundational Ontologies'. In: *Advances in Conceptual Modeling*. Vol. 7518. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 172–181. DOI: `10.1007/978-3-642-33999-8_21`.

[36] John Guerson et al. 'OntoUML Lightweight Editor: A Model-Based Environment to Build, Evaluate and Implement Reference Ontologies'. In: *2015 IEEE 19th International Enterprise Distributed Object Computing Workshop*. 2015, pp. 144–147. DOI: `10.1109/EDOCW.2015.17`.

[37] Monalessa Perini Barcellos and Ricardo De Almeida Falbo. 'Using a Foundational Ontology for Reengineering a Software Enterprise Ontology'. In: *Advances in Conceptual Modeling - Challenging Perspectives*. Ed. by Carlos Alberto Heuser and Günther Pernul. Vol. 5833. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 179–188. DOI: `10.1007/978-3-642-04947-7_22`.

[38] Nan Tang. 'Big RDF data cleaning'. In: *2015 31st IEEE International Conference on Data Engineering Workshops*. Apr. 2015, pp. 77–79. DOI: `10.1109/ICDEW.2015.7129549`.

[39] W3C. *OWL Web Ontology Language Overview*. 2004. URL: `https://www.w3.org/OWL/`.

[40] Simon Jupp, Sean Bechhofer and Robert Stevens. 'SKOS with OWL: Don't be Full-ish!' In: *OWLED*. Vol. 432. 2008, pp. 2009–2. URL: `https://ceur-ws.org/Vol-432/owled2008eu_submission_22.pdf`.

[41] Filipi Miranda Soares et al. 'Building a Community-Based FAIR Metadata Schema for Brazilian Agriculture and Livestock Trading Data'. In: *SEMPDW 2022*. CEUR workshop proceedings. Rheinisch Westfälische Technische Hochschule, 2022.

[42] C. Maria Keet. *An introduction to ontology engineering*. London: College Publications, 2018. ISBN: 9781848902954.

[43] Muhammad Ayaz et al. 'The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities'. en. In: *JMIR Medical Informatics* 9.7 (July 2021), e21929. DOI: `10.2196/21929`.

[44] Carina Nina Vorisek et al. 'Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review'. en. In: *JMIR Medical Informatics* 10.7 (July 2022), e35724. DOI: `10.2196/35724`.

[45] Tessa Ohlsen et al. 'PCEtoFHIR: Decomposition of Postcoordinated SNOMED CT Expressions for Storage as HL7 FHIR Resources'. en. In: 12 (Sept. 2024), e57853. DOI: `10.2196/57853`.

[46] Prabath Jayathissa and Roshan Hewapathrana. 'HAPI-FHIR Server Implementation to Enhancing Interoperability among Primary Care Health Information Systems in Sri Lanka: Review of the Technical Use Case'. In: *European Modern Studies Journal* 7.6 (Feb. 2024), pp. 225–241. DOI: `10.59573/emsj.7(6).2023.23`.

[47] John Wieczorek et al. 'Darwin Core: An Evolving Community-Developed Biodiversity Data Standard'. In: *PLoS ONE* 7.1 (Jan. 2012). Ed. by Indra Neil Sarkar, e29715. DOI: `10.1371/journal.pone.0029715`.

[48] Tim Robertson et al. 'The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet'. en. In: *PLoS ONE* 9.8 (Aug. 2014). Ed. by Damon P. Little, e102623. DOI: `10.1371/journal.pone.0102623`.

[49] Eric H. Fegraus et al. 'Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation'. en. In: *Bulletin of the Ecological Society of America* 86.3 (July 2005), pp. 158–168. DOI: `10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2`.

[50] Ed Baker, Simon Rycroft and Vincent Smith. 'Linking multiple biodiversity informatics platforms with Darwin Core Archives'. In: *Biodiversity Data Journal* 2 (Jan. 2014), e1039. DOI: `10.3897/BDJ.2.e1039`.

[51] Steven J. Baskauf and Campbell O. Webb. 'Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF'. In: *Semantic Web* 7.6 (Oct. 2016). Ed. by Pascal Hitzler et al., pp. 629–643. DOI: `10.3233/SW-150203`.

[52] Pier Buttigieg et al. 'The environment ontology: contextualising biological and biomedical entities'. en. In: *Journal of Biomedical Semantics* 4.1 (2013), p. 43. DOI: `10.1186/2041-1480-4-43`.

[53] Ramona L. Walls et al. 'Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies'. en. In: *PLoS ONE* 9.3 (Mar. 2014). Ed. by Vladimir B. Bajic, e89606. DOI: `10.1371/journal.pone.0089606`.

[54] William Michener et al. 'DataONE: Data Observation Network for Earth Preserving Data and Enabling Innovation in the Biological and Environmental Sciences'. en. In: *D-Lib Magazine* 17.1/2 (Jan. 2011). DOI: `10.1045/january2011-michener`.

[55] J. Frederick Grassle. 'The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context'. In: *Oceanography* 13.3 (2000), pp. 5–7. DOI: `10.5670/oceanog.2000.01`.

[56] Robert Battle and Dave Kolas. 'Enabling the geospatial Semantic Web with Parliament and GeoSPARQL'. In: *Semant. Web* 3.4 (Oct. 2012), pp. 355–370.

[57] Nicholas J. Car and Timo Homburg. 'GeoSPARQL 1.1: Motivations, Details and Applications of the Decadal Update to the Most Important Geospatial LOD Standard'. en. In: *ISPRS International Journal of Geo-Information* 11.2 (Feb. 2022), p. 117. DOI: `10.3390/ijgi11020117`.

[58] Milos Jovanovik, Timo Homburg and Mirko Spasi . 'A GeoSPARQL Compliance Benchmark'. In: *ISPRS International Journal of Geo-Information* 10.7 (2021). DOI: `10.3390/ijgi10070487`.

[59] Fayçal Hamdi et al. 'GeomRDF: A Geodata Converter with a Fine-Grained Structured Representation of Geometry in the Web'. In: *Proceedings of the 1st International Workshop on Geospatial Linked Data (GeoLD 2014) - SEMANTiCS 2014*. Germany, Sept. 2014. URL: `https://dblp.org/rec/journals/corr/HamdiABF15.html`.

[60] European Parliament and Council of the European Union. *Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. Official Journal of the European Union, L 108, 25.4.2007, p. 1–14. Mar. 2007. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32007L0002`.

[61] A Troumpoukis, S Konstantopoulos and N Prokopaki-Kostopoulou. ‘A geospatial source selector for federated GeoSPARQL querying’. In: *Open Research Europe* 2.48 (2022). DOI: `10.12688/openreseurope.14605.2`.

[62] Henrique Santos, James P. McCusker and Deborah L. McGuinness. ‘Geospatial Reasoning with Shapefiles for Supporting Policy Decisions’. In: *Proceedings of the 4th International Workshop on Geospatial Linked Data (GeoLD 2021), co-located with the 18th Extended Semantic Web Conference (ESWC 2021)*. Ed. by Beyza Yaman et al. June 2021, Paper 14. URL: `https://ceur-ws.org/Vol-2977/paper14.pdf`.

[63] Efstratios - Paschalis D. Kakalis. *Advancing GeoSPARQL Query Generation on YAGO2Geo: Leveraging Large Language Models and Automated URI Injection from Natural Language Questions*. BSc Thesis. Supervisors: Manolis Koubarakis, Professor; Sergios-Anestis Kefalidis, Associate Researcher. Athens, Greece, Oct. 2024. URL: `https://pergamos.lib.uoa.gr/uoa/dl/object/3420253`.

[64] Yu Feng, Linfang Ding and Guohui Xiao. ‘GeoQAMap - Geographic Question Answering with Maps Leveraging LLM and Open Knowledge Base’. In: *12th International Conference on Geographic Information Science (GIScience 2023)*. Ed. by Roger Beecham et al. Vol. 277. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023, pp. 1–7. DOI: `10.4230/LIPIcs.GIScience.2023.28`.

[65] P Salma Khatoon and Muqeem Ahmed. ‘Importance of semantic interoperability in smart agriculture systems’. In: *Transactions on Emerging Telecommunications Technologies* 33.5 (May 2022). DOI: `10.1002/ett.4448`.

[66] Phokion G. Kolaitis. ‘Reflections on Schema Mappings, Data Exchange, and Metadata Management’. In: *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. PODS ’18. Houston, TX, USA: Association for Computing Machinery, 2018, pp. 107–109. DOI: `10.1145/3196959.3196991`.

[67] Giancarlo Guizzardi. ‘Ontology, Ontologies and the “I” of FAIR’. In: *Data Intelligence* 2.1-2 (Jan. 2020), pp. 181–191. DOI: `10.1162/dint_a_00040`.

[68] Muhammad Asfand-E-Yar and Ramis Ali. ‘Semantic Integration of Heterogeneous Databases of Same Domain Using Ontology’. In: *IEEE Access* 8 (2020), pp. 77903–77919. DOI: `10.1109/ACCESS.2020.2988685`.

[69] Mohamed A. G. Hazber et al. ‘Integration Mapping Rules: Transforming Relational Database to Semantic Web Ontology’. In: *Applied Mathematics and Information Sciences* 10.3 (May 2016), pp. 881–901. DOI: `10.18576/amis/100307`.

[70] Paul Johannesson and Erik Perjons. *An Introduction to Design Science*. Cham: Springer International Publishing, 2014. DOI: `10.1007/978-3-319-10632-8`.

[71] Roel J. Wieringa. *Design Science Methodology for Information Systems and Software Engineering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. DOI: `10.1007/978-3-662-43839-8`.

[72] Lars Vogt et al. ‘Suggestions for extending the FAIR Principles based on a linguistic perspective on semantic interoperability’. en. In: *Scientific Data* 12.1 (Apr. 2025), p. 688. DOI: `10.1038/s41597-025-05011-x`.

[73] Ricardo de Almeida Falbo. ‘SABiO: Systematic Approach for Building Ontologies’. In: *Proceedings of the 1st Joint Workshop on Ontologies in Conceptual Modeling and Information Systems Engineering*. Ed. by Giancarlo Guizzardi et al. Vol. 1301. CEUR Workshop Proceedings. Rio de Janeiro, Brazil: CEUR-WS.org, 2014. URL: `https://ceur-ws.org/Vol-1301/ontocomodise2014_2.pdf`.

[74] Camila Zachhe Aguiar and Vítor E. Silva Souza. ‘SABiOx: The Extended Systematic Approach for Building Ontologies’. In: *Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024)*. Vitoria, Brazil: CEUR Workshop Proceedings, Oct. 2024. URL: `https://ceur-ws.org/Vol-3905/paper1.pdf`.

[75] Bizagi BPM. *BPMN Guide Quick Start*. 2017. URL: `https://www.bizagi.com/files/live/sites/bizagi/files/PDFs/BPMN_Quick_Start_Guide.pdf`.

[76] Debora Drucker et al. ‘Desdobramentos da Implementação da Rede GO FAIR Agro Brasil no Biênio 2021-2023’. In: *Anais do XIV Congresso Brasileiro de Agroinformática*. Natal/RN: SBC, 2023, pp. 286–293. DOI: `10.5753/sbiagro.2023.26570`.

[77] José A Salim et al. 'Data standardization of plant-pollinator interactions'. In: *GigaScience* 11 (May 2022), giac043. DOI: `10.1093/gigascience/giac043`.

[78] Maarten Trekels et al. *WorldFAIR Project (D10.1) Agriculture-related pollinator data standards use cases report*. WorldFAIR Project, July 2023. DOI: `10.5281/zenodo.8176978`.

[79] Debora Drucker et al. *WorldFAIR (D10.2) Agricultural Biodiversity Standards, Best Practices and Guidelines Recommendations*. Zenodo, Feb. 2024. DOI: `10.5281/ZENODO.10666593`.

[80] Drucker, Debora et al. *WorldFAIR (D10.3) Agricultural biodiversity FAIR data assessment rubrics*. Zenodo, Feb. 2024. DOI: `10.5281/zenodo.10719265`.

[81] Drucker, Debora Pignatari et al. 'Plant-pollinator Interaction Data: A case study of the WorldFAIR project'. In: *Biodiversity Information Science and Standards* 6 (Sept. 2022). DOI: `10.3897/biss.6.94310`.

[82] Hilde Orten. *Cross Domain Integration (CDI) Working Group*. 2024. URL: `https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/2681241718/Cross+Domain+Integration+CDI+Working+Group`.

[83] Bernabé, César H. et al. 'A goal-oriented method for FAIRification planning'. In: *CEUR Workshop Proceedings* 3618 (2023). URL: `https://er2023.inesc-id.pt/`.

[84] Kendall Lister. 'Toward Semantic Interoperability for Software Systems'. Department of Engineering - Computer Science and Software Engineering. PhD thesis. Melbourne, Australia: The University of Melbourne, 2008. URL: `http://hdl.handle.net/11343/35107`.

[85] Mazhar Hameed and Felix Naumann. 'Data Preparation: A Survey of Commercial Tools'. In: *ACM SIGMOD Record* 49.3 (Dec. 2020), pp. 18–29. DOI: `10.1145/3444831.3444835`.

[86] Tamraparni Dasu and Theodore Johnson. *Exploratory Data Mining and Data Cleaning*. 1st ed. USA: John Wiley & Sons, Inc., 2003. ISBN: 0471268518.

[87] Sean Kandel et al. 'Research directions in data wrangling: Visualizations and transformations for usable and credible data'. In: *Information Visualization* 10.4 (2011), pp. 271–288. DOI: `10.1177/1473871611415994`.

[88] Yongxin Liao et al. 'Why, Where and How to use Semantic Annotation for Systems Interoperability'. In: *Proceedings of the 1st UNITE Doctoral Symposium*. June 2011.

[89] Sanju Tiwari, Fernando Ortiz-Rodriguez and M.A. Jabbar. 'Semantic modeling for healthcare applications: an introduction'. In: *Semantic Models in IoT and eHealth Applications*. Elsevier, 2022, pp. 1–17. DOI: `10.1016/B978-0-32-391773-5.00007-8`.

[90] Yun Lin et al. 'Semantic Annotation Framework to Manage Semantic Heterogeneity of Process Models'. In: *Advanced Information Systems Engineering*. Ed. by Eric Dubois and Klaus Pohl. Berlin, Heidelberg: Springer, 2006, pp. 433–446. DOI: `10.1007/11767138_29`.

[91] E. Amdouni and C. Jonquet. 'FAIR or FAIRer? An Integrated Quantitative FAIRness Assessment Grid for Semantic Resources and Ontologies'. In: *Metadata and Semantic Research. MTSR 2021*. Ed. by E. Garoufallou, M. A. Ovalle-Perandones and A. Vlachidis. Vol. 1537. Communications in Computer and Information Science. Cham: Springer, 2022. DOI: `10.1007/978-3-030-98876-0_6`.

[92] Dag I. K. Sjøberg and Gunnar Rye Bergersen. 'Construct Validity in Software Engineering'. In: *IEEE Transactions on Software Engineering* 49.3 (2023), pp. 1374–1396. DOI: `10.1109/TSE.2022.3176725`.

[93] Filipi Miranda Soares et al. 'Towards a Conceptual Model for FAIR Metadata Schemas'. In: *ER2024: Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos*. Pittsburgh, PA, 2024. URL: `https://ceur-ws.org/Vol-3849/forum4.pdf`.

[94] Yann Le Franc et al. *D2.2 FAIR Semantics: First recommendations*. Mar. 2020. DOI: `10.5281/ZENODO.5361930`.

[95] M. Poveda-Villalón et al. 'Coming to Terms with FAIR Ontologies'. In: *Knowledge Engineering and Knowledge Management. EKAW 2020*. Ed. by C. M. Keet and M. Dumontier. Vol. 12387. Lecture Notes in Computer Science. Cham: Springer, 2020. DOI: `10.1007/978-3-030-61244-3_18`.

[96] Gary Berg-Cross and Sage Arbor. 'Beyond Simple FAIR Principles for Ontologies and Semantic Resources: Grounding Rich, Meaningful Metadata'. In: *Journal of the Washington Academy of Sciences* 108.4 (2022), pp. 1–26. URL: `https://www.jstor.org/stable/27281297`.

[97] Giancarlo Guizzardi et al. 'Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model'. In: *Conceptual Modeling*. Ed. by Aditya Ghose et al. Cham: Springer International Publishing, 2021, pp. 356–366. DOI: `10.1007/978-3-030-89022-3_28`.

[98] Marcia Lei Zeng and Jian Qin. *Metadata*. Third edition. Chicago: ALA Neal-Schuman, 2022. ISBN: 9780838937969.

[99] Lois Mai Chan and Marcia Lei Zeng. 'Metadata Interoperability and Standardization - A Study of Methodology Part I: Achieving Interoperability at the Schema Level'. In: *D-Lib Magazine* 12.6 (June 2006). DOI: `10.1045/june2006-chan`.

[100] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 11179-1:2023: Information technology: Metadata registries (MDR): Part 1: Framework*. Tech. rep. ISO/IEC, 2023.

[101] International Organization for Standardization. *Building a Metadata Schema – Where to Start*. Geneva: ISO, 2008.

[102] International Organization for Standardization and International Electrotechnical Commission. *Information technology - Metadata registries (MDR) - Part 4: Formulation of data definitions*. Tech. rep. 11179-4. ISO/IEC, 2004.

[103] International Organization for Standardization and International Electrotechnical Commission. *Information technology - Metadata registries (MDR) - Part 5: Naming principles*. Tech. rep. 11179-5. ISO/IEC, 2015.

[104] Cross-Domain Interoperability Framework (CDIF) Working Group et al. *Cross Domain Interoperability Framework (CDIF): Discovery Module (v01 draft for public consultation)*. Dec. 2023. DOI: `10.5281/ZENODO.10252564`.

[105] Pierre-Yves Vandenbussche et al. 'Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web'. In: *Semantic Web* 8.3 (Jan. 2017), pp. 437–452. DOI: `10.3233/SW-160213`.

[106] Edison Ong et al. 'Ontobee: A linked ontology data server to support ontology term dereferencing, linkage, query and integration'. In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D347–D352. DOI: `10.1093/nar/gkw918`.

[107] Mark A. Musen, Martin J. O'Connor, Erik Schultes et al. 'Modeling community standards for metadata as templates makes data FAIR'. In: *Scientific Data* 9.1 (2022), p. 696. DOI: `10.1038/s41597-022-01815-3`.

[108] Robert L Hurt et al. 'Sharing images intelligently: The Astronomy Vizualisation Metadata standard'. In: *CAP'07: Proceedings of the Conference on Communicating Astronomy with the Public. Eugenides Foundation*. 2007, pp. 450–453. URL: `https://ui.adsabs.harvard.edu/abs/2008ca07.conf..450H/abstract`.

[109] Mary Vardigan, Pascal Heus and Wendy Thomas. 'Data Documentation Initiative: Toward a standard for the social sciences'. In: *International Journal of Digital Curation* 3.1 (2008). DOI: `10.2218/ijdc.v3i1.45`.

[110] Sarah Higgins. *What are Metadata Standards*. 2007. URL: `https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/what-are-metadata-standards`.

[111] Fulvio Mazzocchi. 'Knowledge Organization System (KOS): An Introductory Critical Account'. In: *KNOWLEDGE ORGANIZATION* 45.1 (2018), pp. 54–78. DOI: `10.5771/0943-7444-2018-1-54`.

[112] Gail M Hodge. *Systems of knowledge organization for digital libraries: beyond traditional authority files*. Vol. 91. Digital Library Federation, 2000. URL: `https://www.clir.org/wp-content/uploads/sites/6/pub91.pdf`.

[113] Jenn Riley. *Understanding metadata: What is Metadata, and What is it For?: A Primer*. Washington, DC: National Information Standards Organization, 2017, pp. 7–10. URL: `https://www.niso.org/publications/understanding-metadata-2017`.

[114] Joan M. Reitz. *Online Dictionary for Library and Information Science*. 2023. URL: `https://odlis.abc-clio.com/odlis_c.html`.

[115] Dan Brickley and R.V. Guha. *RDF Schema 1.1*. W3C Recommendation, Document Status Update: December 1, 2023. Previous editors: Brian McBride. Feb. 2014. URL: `http://www.w3.org/TR/rdf11-schema/`.

[116] Pance Panov, Larisa N. Soldatova and Saso Dzeroski. 'Generic ontology of datatypes'. In: *Information Sciences* 329 (2016). Special issue on Discovery Science, pp. 900–920. DOI: `10.1016/j.ins.2015.08.006`.

[117] Dublin Core Metadata Initiative. *Domains and Ranges*. 2024. URL: `https://www.dublincore.org/resources/glossary/domains_and_ranges/`.

[118] RDF Working Group. *Resource Description Framework (RDF)*. 2014. URL: `https://www.w3.org/RDF/`.

[119] W3C OWL Working Group. *Web Ontology Language (OWL)*. 2012. URL: `https://www.w3.org/OWL/`.

[120] Food and Agriculture Organization of the United Nations. *The State of Agricultural Commodity Markets 2022*. FAO, June 2022. DOI: `10.4060/cc0471en`.

[121] Jiayu Zhou et al. 'On Building Real Time Intelligent Agricultural Commodity Trading Models'. In: *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2022, pp. 89–95. DOI: `10.1109/BigDataService55688.2022.00021`.

[122] Center for Advanced Studies on Applied Economics (CEPEA). *Cepea Agricultural Prices*. 2024. URL: `https://cepea.esalq.usp.br/en`.

[123] F. Soares, F. E. Corrêa and Centro de Estudos Avançados em Economia Aplicada (CEPEA). *Raw Data from Cepea on Sugar, Fed Cattle, Coffee, and Soybean Price Indexes*. dataset. 2024. DOI: `10.5281/zenodo.12163228`.

[124] Instituto de Pesquisa Econômica Aplicada (IPEA). *IPEAdata: indicadores macroeconomicos*. 2024. URL: `http://www.ipeadata.gov.br/Default.aspx`.

[125] F. Soares, F. E. Corrêa and Instituto de Pesquisa Econômica Aplicada (Ipea). *Raw Price Index Data from IpeaData on Leather and Leather Goods, Cellulose Pulp, Paper and Paper Products, Tobacco Products, and Wood Products*. dataset. 2024. DOI: `10.5281/zenodo.12169699`.

[126] Companhia Nacional de Abastecimento (CONAB). *CONAB - Preços Agropecuários*. 2024. URL: `https://www.conab.gov.br/info-agro/precos`.

[127] F. Soares, F. E. Corrêa and Companhia Nacional de Abastecimento (Conab). *Raw Price Index Data from Conab on Sugar, Fed Cattle, Coffee, and Soybean*. dataset. 2024. DOI: `10.5281/zenodo.12170310`.

[128] F. Soares. *Treated Price Index Data from Cepea, Ipea, and Conab*. dataset. 2024. DOI: `10.5281/zenodo.12531081`.

[129] Filipi Miranda Soares et al. 'Unveiling Knowledge Organization Systems' Artifacts for Digital Agriculture with Lexical Network Analysis'. In: *Advances in Conceptual Modeling*. Ed. by Tiago Prince Sales et al. Cham: Springer Nature Switzerland, 2023, pp. 299–311. DOI: `10.1007/978-3-031-47112-4_28`.

[130] EJP RD Virtual Platform. *EJP RD Metadata Schema Classes*. 2024. URL: `https://vp-onboarding-doc.readthedocs.io/en/stable/level_1/classes.html`.

[131] F. Soares et al. *Agriculture and Livestock Metadata Elements Set (Almes Core)*. Data set. 2024. DOI: `10.5281/zenodo.12711290`.

[132] Milena Ambrosio Telles et al. 'Embrapa's Contributions to Integrate Brazilian Agricultural Vocabularies: Agrotermos in AgroPortal'. In: *Proceedings of the 17th Seminar on Ontology Research in Brazil (ONTOBRAS 2024) and 8th Doctoral and Masters Consortium on Ontologies (WTDO 2024)*. Vol. 3905. Vitória, Brazil: CEUR Workshop Proceedings, 2024. URL: `https://ceur-ws.org/Vol-3905/short8.pdf`.

[133] W3C Permanent Identifier Community Group. *W3ID: Permanent Identifiers for the Web*. 2024. URL: `https://w3id.org/`.

[134] Dublin Core Collection Description Task Group. *Dublin Core Collection Description Frequency Vocabulary*. June 2013. URL: `https://dublincore.org/specifications/dublin-core/collection-description/frequency/`.

[135] Receita Federal do Brasil. *NCM - Nomenclatura Comum do Mercosul*. 2024. URL: `https://www.gov.br/receitafederal/pt-br/assuntos/aduana-e-comercio-exterior/classificacao-fiscal-de-mercadorias/download-ncm-nomenclatura-comum-do-mercosul`.

[136] United Nations Statistics Division. *Economic Classifications*. 2024. URL: `https://unstats.un.org/unsd/classifications/Econ`.

[137] Velde-Koerts T van der, Muller E and Ossendorp BHC. *Classification of crops grown in or imported into the European Union for pesticide residue assessment*. 2003. URL: `http://hdl.handle.net/10029/260672`.

[138] European Food Safety Authority (EFSA) et al. 'The 2022 European Union report on pesticide residues in food'. In: *EFSA Journal* 22.4 (2024), e8753. DOI: `10.2903/j.efsa.2024.8753`.

[139] Ahmad Khalid Sadid. 'Scientific and Agronomic Classification of Field Plants'. In: *International Journal for Research in Applied Sciences and Biotechnology* 9.2 (Mar. 2022), pp. 7–10. DOI: `10.31033/ijrasb.9.2.2`.

[140] Sila Maria Rodrigues Ferreira. 'Classificação de Produtos de Origem Vegetal'. In: *Visão Acadêmica* 6.1 (June 2005). DOI: `10.5380/acd.v6i1.572`.

[141] Madhur Gautam and Rashid Faruqee. 'Connecting Farmers to Markets: Trading in High-Value Products'. In: *Dynamics of Rural Growth in Bangladesh: Sustaining Poverty Reduction*. Washington, D.C.: World Bank Publications, June 2016, pp. 145–174. DOI: `10.1596/978-1-4648-0876-0_ch5`.

[142] Michelle Márcia Viana Martins. ‘Efeitos da regulamentação sobre o comércio de produtos agroalimentares’. Doutorado em Economia Aplicada. Piracicaba: Universidade de São Paulo, July 2021. DOI: `10.11606/T.11.2021.tde-10092021-133022`.

[143] Clément Jonquet et al. ‘AgroPortal: A vocabulary and ontology repository for agronomy’. In: *Computers and Electronics in Agriculture* 144 (Jan. 2018), pp. 126–143. DOI: `10.1016/j.compag.2017.10.012`.

[144] Jim Reecy and Cari Park. *Livestock Product Trait Ontology (LPT)*. Dec. 2022. DOI: `10.25504/FAIRsharing.g78mbm`.

[145] Baptiste Darnala et al. ‘C3PO: a crop planning and production process ontology and knowledge graph’. In: *Frontiers in Artificial Intelligence* 6 (Oct. 2023). DOI: `10.3389/frai.2023.1187090`.

[146] Moise Kombolo et al. *Thésaurus de la viande*. Version V1. 2022. DOI: `10.15454/PB5QXC`.

[147] Liliana Andres-Hernandez et al. ‘Establishing a Common Nutritional Vocabulary - From Food Production to Diet’. In: *Frontiers in Nutrition* 9 (June 2022), p. 928837. DOI: `10.3389/fnut.2022.928837`.

[148] A. R. Muljarto et al. ‘Ontology-Based Model for Food Transformation Processes - Application to Winemaking’. In: *Metadata and Semantics Research*. Ed. by S. Closs et al. Vol. 478. Communications in Computer and Information Science. Springer, Cham, 2014, pp. 378–389. DOI: `10.1007/978-3-319-13674-5_30`.

[149] Damion M. Dooley et al. ‘FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration’. In: *npj Science of Food* 2.1 (2018), p. 23. DOI: `10.1038/s41538-018-0032-6`.

[150] Food and Agriculture Organization of the United Nations. *Embrapa: AGROVOC and GTermos*. 2022. URL: `https://www.fao.org/agrovoc/news/embrapa-agrovoc-and-gtermos`.

[151] Emna Amdouni, Syphax Bouazzouni and Clement Jonquet. ‘O’faire: ontology fairness evaluator in the agroportal semantic resource repository’. In: *The Semantic Web: ESWC 2022 Satellite Events*. Ed. by Paul Groth et al. Vol. 13384. Cham: Springer International Publishing, 2022, pp. 89–94. DOI: `10.1007/978-3-031-11609-4_17`.

[152] Mark A. Musen. ‘The protégé project: a look back and a look forward’. In: *AI Matters* 1.4 (June 2015), pp. 4–12. DOI: `10.1145/2757001.2757003`.

[153] Filipi Miranda Soares et al. *Agriculture Product Types Ontology (APTO)*. Zenodo, Aug. 2024. DOI: `10.5281/zenodo.13271285`.

[154] O Estado de S. Paulo. *Qual a diferença entre leite, bebida láctea e soro de leite?* 2022. URL: `https://www.estadao.com.br/economia/diferenca-leite-bebida-lactea-soro-de-leite/`.

[155] International Commission on Zoological Nomenclature. *International Code of Zoological Nomenclature*. Ed. by W.D.L. Ride et al. Fourth. London: The International Trust for Zoological Nomenclature, 1999. ISBN: 0 85301 006 4.

[156] GBIF Secretariat. *Prochilodus cearensis Steindachner, 1911 in GBIF Secretariat (2023)*. Checklist dataset accessed via GBIF.org on 2024-08-15; GBIF Backbone Taxonomy. 2023. DOI: `10.15468/39omei`. URL: `https://www.gbif.org`.

[157] W.D.L. Ride et al., eds. *International Code of Zoological Nomenclature*. 4th ed. London: The International Trust for Zoological Nomenclature, 2012. ISBN: 0 85301 006 4. URL: `https://www.iczn.org/the-code/the-code-online/`.

[158] N. J. Turland et al., eds. *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code)*. Vol. 159. Regnum Vegetabile. Adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Glashütten: Koeltz Botanical Books, 2018. DOI: `10.12705/Code.2018`.

[159] Embrapa. *Agrotermos*. 2024. URL: `https://sistemas.sede.embrapa.br/agrotermos/`.

[160] Filipi Miranda Soares et al. 'Exploring a Large Language Model for Transforming Taxonomic Data into OWL: Lessons Learned and Implications for Ontology Development'. en. In: *Data Intelligence* (Mar. 2025). DOI: `10.3724/2096-7004.di.2025.0020`.

[161] Sun Yu, Xia Youming and Zhiping Li. 'Formal contexts in ontologies'. In: *2011 6th International Conference on Computer Science and Education (ICCSE)*. 2011, pp. 214–217. DOI: `10.1109/ICCSE.2011.6028620`.

[162] Jain Sarika and Mishra Sanju. 'Knowledge Representation with Ontology Tools and Methodology'. In: *International Journal of Computer Applications* 107.18 (2014), pp. 1–5. DOI: `10.5120/18841-0042`.

[163] Vânia M. P. Vidal et al. 'An Ontology-Based Framework for Heterogeneous Data Sources Integration'. In: *Revista de Informática Teórica e Aplicada* 16.2 (Mar. 2010), pp. 61–64. DOI: `10.22456/2175-2745.12570`.

[164] Enesi Femi Aminu et al. 'A Review on Ontology Development Methodologies for Developing Ontological Knowledge Representation Systems for various Domains'. In: *International Journal of Information Engineering and Electronic Business* 12.2 (Apr. 2020), pp. 28–39. DOI: `10.5815/ijieeb.2020.02.05`.

[165] Maria Luiza De Almeida Campos et al. 'Methodological aspects concerning the reuse of ontologies: a study based on genomic annotations in the domain of trypanosomatides'. In: *RECIIS* 3.1 (Mar. 2009), pp. 243–257. DOI: `10.3395/reciis.v3i1.243en`.

[166] Miguel E. Ruiz et al. 'Ontological Research and Its Applications to the Biomedical Domain'. In: *Proceedings of the American Society for Information Science and Technology* 42.1 (Jan. 2006), pp. 1-10. DOI: `10.1002/meet.14504201198`.

[167] Cedric Gillott. 'Taxonomy and Systematics'. In: *Entomology*. Dordrecht: Springer Netherlands, 1995, pp. 91–112. DOI: `10.1007/978-94-017-4380-8_4`.

[168] Lorenzo Peruzzi. 'Advances in Plant Taxonomy and Systematics'. In: *Biology* 12.4 (Apr. 2023), p. 570. DOI: `10.3390/biology12040570`.

[169] Matjaž Kuntner and Ingi Agnarsson. 'Are the Linnean and Phylogenetic Nomenclatural Systems Combinable? Recommendations for Biological Nomenclature'. In: *Systematic Biology* 55.5 (Oct. 2006). Ed. by Adrian Paterson, pp. 774–784. DOI: `10.1080/10635150600981596`.

[170] Germinal Rouhan and Myriam Gaudeul. 'Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives'. In: *Molecular Plant Taxonomy*. Ed. by Pascale Besse. Vol. 2222. New York, NY: Springer US, 2021, pp. 1–38. DOI: `10.1007/978-1-0716-0997-2_1`.

[171] Food and Agriculture Organization of the United Nations. *AGROVOC 4: AGROVOC editorial guidelines*. FAO, Feb. 2023. ISBN: 9789251374603. DOI: `10.4060/cc0647en`.

[172] A. Sehar et al. 'Streamlining FoodOn Seafood Nomenclature using a Semi-automated ROBOT Template-driven Approach'. In: *Proceedings of the International Food Ontology Workshop (IFOW)*. 2021. URL: `https://ceur-ws.org/Vol-2969/paper9-IFOW.pdf`.

[173] Xiaohan Feng, Xixin Wu and Helen Meng. 'Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata Schema'. In: *Proceedings of the KDD Workshop on Human-Interpretable AI 2024 (HI-AI 2024), co-located with 30th ACM SIGKDD Conference on Knowledge Discovery*

and Data Mining (KDD 2024). Vol. 3841. CEUR Workshop Proceedings. Centre de Convencions Internacional de Barcelona, Spain: CEUR-WS.org, Aug. 2024, pp. 117–135. URL: `https://ceur-ws.org/Vol-3841/Paper19.pdf`.

[174] Mohammad Javad Saeedizade and Eva Blomqvist. 'Navigating Ontology Development with Large Language Models'. In: *The Semantic Web*. Ed. by Albert Meroño Peñuela et al. Vol. 14664. Cham: Springer Nature Switzerland, 2024, pp. 143–161. DOI: `10.1007/978-3-031-60626-7_8`.

[175] Tanay Aggarwal et al. 'Large Language Models for Scholarly Ontology Generation: An Extensive Analysis in the Engineering Field'. In: *arXiv* (2024). DOI: `10.48550/ARXIV.2412.08258`.

[176] Antonio Zaitoun et al. 'Can Large Language Models Augment a Biomedical Ontology with missing Concepts and Relations?' In: *arXiv* (2023). DOI: `10.48550/ARXIV.2311.06858`.

[177] Milena Trajanoska, Riste Stojanov and Dimitar Trajanov. 'Enhancing Knowledge Graph Construction Using Large Language Models'. In: *arXiv* (2023). DOI: `10.48550/ARXIV.2305.04676`.

[178] Julio Cesar dos Reis and André Gomes Regino. 'Generating E-commerce Related Knowledge Graph from Text: Open Challenges and Early Results using LLMs'. In: *Joint Proceedings of the 3rd International Workshop on Knowledge Graph Generation from Text (TEXT2KG) and Data Quality meets Machine Learning and Knowledge Graphs (DQMLKG)*. Vol. 3747. CEUR Workshop Proceedings. Hersonissos, Greece: CEUR-WS.org, May 2024. URL: `https://ceur-ws.org/Vol-3747/text2kg%5C_paper11.pdf`.

[179] Reihaneh Amini et al. 'Towards Complex Ontology Alignment using Large Language Models'. In: *arXiv* (2024). DOI: `10.48550/ARXIV.2404.10329`.

[180] Nicolas Matentzoglu et al. 'MapperGPT: Large Language Models for Linking and Mapping Entities'. In: *arXiv* (2023). DOI: `10.48550/ARXIV.2310.03666`.

[181] Guanchen Wu et al. 'Ontology extension by online clustering with large language model agents'. In: *Frontiers in Big Data* 7 (Oct. 2024), p. 1463543. DOI: `10.3389/fdata.2024.1463543`.

[182] Levi van Hees. 'Exploring the utility of large language models for achieving semantic interoperability in data ecosystems'. Master Thesis. Utrecht University, 2024. URL: `https://studenttheses.uu.nl/handle/20.500.12932/48039`.

[183] Yunyi Zhang et al. ‘Automated Mining of Structured Knowledge from Text in the Era of Large Language Models’. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD ’24. Barcelona, Spain: Association for Computing Machinery, 2024, pp. 6644–6654. DOI: `10.1145/3637528.3671469`.

[184] Marvin Hofer et al. ‘Construction of Knowledge Graphs: Current State and Challenges’. In: *Information* 15.8 (Aug. 2024), p. 509. DOI: `10.3390/info15080509`.

[185] Jiaoyan Chen et al. ‘Knowledge Graphs for the Life Sciences: Recent Developments, Challenges and Opportunities’. In: *Transactions on Graph Data and Knowledge* 1.1 (2023), pp. 1–33. DOI: `10.4230/TGDK.1.1.5`.

[186] S. Tsaneva, S. Vasic and M. Sabou. ‘LLM-driven Ontology Evaluation: Verifying Ontology Restrictions with ChatGPT’. In: *Data Quality meets Machine Learning and Knowledge Graphs: DQM-LKG Workshop at ESWC 2024*. In press. 2024.

[187] Olga Perera and Jun Liu. ‘Exploring large language models for ontology learning’. In: *Issues in Information Systems* 25.4 (2024), pp. 299–310. DOI: `10.48009/4_iis_2024_124`.

[188] Ozan Baris Mulayim et al. ‘Large Language Models for the Creation and Use of Semantic Ontologies in Buildings: Requirements and Challenges’. In: *Proceedings of the 11th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. BuildSys ’24. Hangzhou, China: Association for Computing Machinery, 2024, pp. 312–317. DOI: `10.1145/3671127.3698792`.

[189] Fabian Neuhaus. ‘Ontologies in the era of large language models – a perspective’. In: *Applied Ontology* 18.4 (Dec. 2023), pp. 399–407. DOI: `10.3233/AO-230072`.

[190] Markus Krötzsch. ‘Ontologies for Knowledge Graphs?’ In: *Proceedings of the 30th International Workshop on Description Logics (DL 2017)*. Ed. by Alessandro Artale, Birte Glimm and Roman Kontchakov. Vol. 1879. CEUR Workshop Proceedings. Montpellier, France: CEUR-WS.org, July 2017. URL: `https://ceur-ws.org/Vol-879/invited2.pdf`.

[191] Markus Krötzsch and Veronika Thost. ‘Ontologies for Knowledge Graphs: Breaking the Rules’. In: *The Semantic Web – ISWC 2016*. Ed. by Paul Groth et al. Vol. 9981. Cham: Springer International Publishing, 2016, pp. 376–392. DOI: `10.1007/978-3-319-46523-4_23`.

[192] Nandana Mihindukulasooriya et al. 'Text2KGBench: A Benchmark for Ontology-Driven Knowledge Graph Generation from Text'. In: *The Semantic Web – ISWC 2023*. Ed. by Terry R. Payne et al. Vol. 14266. Cham: Springer Nature Switzerland, 2023, pp. 247–265. DOI: `10.1007/978-3-031-47243-5_14`.

[193] Aidan Hogan et al. 'Knowledge graphs'. In: *ACM Comput. Surv.* 54.4 (May 2022), pp. 1–37.

[194] Abhijeet Kumar et al. 'Building Knowledge Graph using Pre-trained Language Model for Learning Entity-aware Relationships'. In: *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON)*. 2020, pp. 310–315. DOI: `10.1109/GUCON48875.2020.9231227`.

[195] Antoine Bosselut et al. 'COMET: Commonsense Transformers for Automatic Knowledge Graph Construction'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019, pp. 4762–4779. DOI: `10.18653/v1/P19-1470`.

[196] Global Biodiversity Information Facility. *GBIF Home Page*. 2024. URL: `https://www.gbif.org/`.

[197] Beckett Sterner et al. 'Unified and pluralistic ideals for data sharing and reuse in biodiversity'. In: *Database* 2023 (July 2023), baad048. DOI: `10.1093/database/baad048`.

[198] Convention on Biological Diversity and Global Biodiversity Information Facility. *Big data for biodiversity: Global Biodiversity Information Facility surpasses one billion records*. 2018. URL: `https://www.cbd.int/doc/press/2018/pr-2018-07-09-cbd-gbif-en.pdf`.

[199] GBIF Secretariat. *GBIF Backbone Taxonomy*. Accessed via GBIF.org on 2024-03-04. 2023. URL: `https://doi.org10.15468/39omei`.

[200] GBIF Secretariat. *GBIF API Reference*. Accessed via GBIF.org on 2024-03-04. URL: `https://techdocs.gbif.org/en/openapi/`.

[201] Donald Hobern et al. 'Towards a global list of accepted species VI: The Catalogue of Life checklist'. In: *Organisms Diversity and Evolution* 21.4 (Dec. 2021), pp. 677–690. DOI: `10.1007/s13127-021-00516-w`.

[202] W. Greuter et al. 'Draft BioCode (1997): The Prospective International Rules for the Scientific Names of Organisms'. In: *Taxon* 47.1 (1998), pp. 127–150. DOI: `10.2307/1224030`.

[203] Judith E Winston. ‘Twenty-First Century Biological Nomenclature—The Enduring Power of Names’. In: *Integrative and Comparative Biology* 58.6 (Sept. 2018), pp. 1122–1131. DOI: `10.1093/icb/icy060`.

[204] Samir Okasha. ‘Species and classification’. In: *Philosophy of Biology: A Very Short Introduction*. Oxford University Press, Nov. 2019. DOI: `10.1093/actrade/9780198806998.003.0005`.

[205] Louie Giray. ‘Prompt Engineering with ChatGPT: A Guide for Academic Writers’. In: *Annals of Biomedical Engineering* 51.12 (Dec. 2023), pp. 2629–2633. DOI: `10.1007/s10439-023-03272-4`.

[206] OpenAI. *Meet BrowserOp: our ChatGPT plugin that lets you search up-to-date information smoothly and efficiently*. 2023. URL: `https://community.openai.com/t/meet-browserop-our-chatgpt-plugin-that-lets-you-search-up-to-date-information-smoothly-and-efficiently/304018`.

[207] Zekun Wang et al. *Interactive Natural Language Processing*. 2023. DOI: `10.48550/ARXIV.2305.13246`.

[208] Tatwadarshi P. Nagarhalli et al. ‘The Review of Natural Language Processing Applications with Emphasis on Machine Learning Implementations’. In: *2022 International Conference on Electronics and Renewable Systems (ICEARS)*. Tuticorin, India: IEEE, Mar. 2022, pp. 1353–1358. DOI: `10.1109/ICEARS53579.2022.9752326`.

[209] Thanveer Shaik et al. ‘A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis’. In: *IEEE Access* 10 (2022), pp. 56720–56739. DOI: `10.1109/ACCESS.2022.3177752`.

[210] Okan Bulut and Tarid Wongvorachan. ‘Feedback Generation through Artificial Intelligence’. In: *The Open/Technology in Education, Society, and Scholarship Association Conference* 2.1 (Dec. 2022), pp. 1–9. DOI: `10.18357/otessac.2022.2.1.125`.

[211] Kew Royal Botanic Gardens. *Plants of the World Online*. 2024. URL: `http://www.plantsoftheworldonline.org/`.

[212] N. J. Turland et al. ‘Chapter H: Names of Hybrids’. In: *International Code of Nomenclature for Algae, Fungi, and Plants (Shenzhen Code)*. Koeltz Botanical Books, 2019. URL: `https://www.iapt-taxon.org/nomen/pages/main/art_h3.html`.

[213] Food and Agriculture Organization of the United Nations. ‘Pepper’. In: *Agrovoc*. Food and Agriculture Organization of the United Nations, 2011. URL: `http://aims.fao.org/aos/agrovoc/c_25495`.

[214] Embrapa. 'Pimenta'. In: *Agrotermos: Espaço conceitual da Embrapa*. Embrapa, 2017. URL: `https://sistemas.sede.embrapa.br/agrotermos/resources/bcf87c03ae11d3d7a749a75c87b4649a`.

[215] Food and Agriculture Organization of the United Nations. 'Chillies'. In: *Agrovoc*. FAO, 2012. URL: `http://aims.fao.org/aos/agrovoc/c_9795`.

[216] Embrapa. 'Condimento picante'. In: *Agrotermos: Espaço conceitual da Embrapa*. Embrapa, 2017. URL: `https://sistemas.sede.embrapa.br/agrotermos/resources/22538a47e4cb9e2ca6a5ed5d3fc59ce3`.

[217] Brasil. Ministério da Saúde. Secretaria de Políticas de Saúde. Coordenação-Geral da Política de Alimentação e Nutrição. *Alimentos regionais brasileiros*. Série F. Comunicação e Educação em Saúde 21. Brasília: Ministério da Saúde, 2002, pp. 1–140.

[218] Gabrielli Menezes. *Cheiro-verde é coentro? Entenda a combinação de temperos*. Sept. 2021. URL: `https://www.uol.com.br/nossa/noticias/redacao/2021/09/17/cheiro-verde-e-coentro-entenda-o-que-e-o-tempero.htm?cmpid=copiaecola`.

[219] Franz Eugen Köhler. 'Petroselinum crispum'. In: *Koehler's Medizinal-Pflanzen*. Wikimedia Commons, 1897. URL: `https://commons.wikimedia.org/wiki/File:Petroselinum_crispum_-_K%C3%B6hler%E2%80%93s_Medizinal-Pflanzen-103.jpg`.

[220] Franz Eugen Köhler. 'Coriandrum sativum'. In: *Koehler's Medizinal-Pflanzen*. Wikimedia Commons, 1897. URL: `https://commons.wikimedia.org/wiki/File:Coriandrum_sativum_-_K%C3%B6hler%E2%80%93s_Medizinal-Pflanzen-193.jpg`.

[221] Carl Axel Magnus Lindman. 'Allium schoenoprasum L.' In: *Bilder ur Nordens Flora*. Stockholm: Wikimedia Commons, 1926. URL: `https://commons.wikimedia.org/wiki/File:386_Allium_schoenoprasum.jpg`.

[222] DXLINH. *Cayngotau*. 2022. URL: `https://commons.wikimedia.org/wiki/File:Cayngotau.jpg`.

[223] Martijanti Martijanti, Sutarno Sutarno and Ariadne L. Juwono. 'Polymer Composite Fabrication Reinforced with Bamboo Fiber for Particle Board Product Raw Material Application'. In: *Polymers* 13.24 (2021). DOI: `10.3390/polym13244377`.

[224] Siti Nikmatin et al. 'Oil Palm Empty Fruit Bunches as Raw Material of Dissolving Pulp for Viscose Rayon Fiber in Making Textile Products'. In: *Polymers* 14.15 (2022). DOI: `10.3390/polym14153208`.

[225] Filipi Miranda Soares et al. 'A Novel UX-Based Approach for Ontology Evaluation: Applying Tree Testing to the Agricultural Product Types Ontology'. In: *IEEE Access* 13 (2025). DOI: `10.1109/ACCESS.2025.3595447`.

[226] Núria Casellas. 'Ontology Evaluation through Usability Measures'. In: *On the Move to Meaningful Internet Systems: OTM 2009 Workshops*. Ed. by Robert Meersman, Pilar Herrero and Tharam Dillon. Vol. 5872. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 594–603. DOI: `10.1007/978-3-642-05290-3_73`.

[227] Bill Albert and Tom Tullis. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*. Newnes, 2013.

[228] Bo Fu, Natalya F. Noy and Margaret-Anne Storey. 'Eye tracking the user experience – An evaluation of ontology visualization techniques'. In: *Semantic Web* 8.1 (Nov. 2016). Ed. by Aba-Sah Dadzie and Emmanuel Pietriga, pp. 23–41. DOI: `10.3233/SW-140163`.

[229] Juan García et al. 'Usability Evaluation of a Visual Modelling Tool for OWL Ontologies'. In: *Journal of Universal Computer Science* 17.9 (2011), pp. 1299–1313. URL: `https://gredos.usal.es/handle/10366/121374`.

[230] Elena García-Barriocanal, Miguel A Sicilia and Salvador Sánchez-Alonso. 'Usability evaluation of ontology editors'. In: *Knowledge Organization* 32.1 (2005), pp. 1–9. DOI: `10.5771/0943-7444-2005-1-1`.

[231] Zeeshan Haider Malik. 'Usability evaluation of ontology engineering tools'. In: *2017 Computing Conference*. London: IEEE, July 2017, pp. 576–584. DOI: `10.1109/SAI.2017.8252154`.

[232] Jinie Pak and Lina Zhou. 'A Framework for Ontology Evaluation'. In: *Exploring the Grand Challenges for Next Generation E-Business*. Ed. by Raj Sharman, H. Raghav Rao and T. S. Raghu. Vol. 52. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 10–18. DOI: `10.1007/978-3-642-17449-0_2`.

[233] Don Norman. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, 2013.

[234] Ezra Schwartz. *Exploring experience design: fusing business, tech, and design to shape customer engagement*. 1st ed. Erscheinungsort nicht ermittelbar: Packt Publishing, 2017. ISBN: 9781787122444.

[235] Jakob Nielsen. *How Many Test Users in a Usability Study?* 2012. URL: `https://www.nngroup.com/articles/how-many-test-users/`.

[236] Optimal Workshop. *How to Interpret the Task Results Tab in Tree Testing*. Learn how to interpret the different kinds of task scores in tree testing. 2024. URL: `https://support.optimalworkshop.com/en/articles/2626846-how-to-interpret-the-task-results-tab-in-tree-testing`.

[237] Luís Antônio Esmerino and Fernanda Rogenski Penteado. 'Avaliação da qualidade microbiológica da carne de frango comercializada no município de Ponta Grossa - Paraná'. In: *Publicatio UEPG: Ciências Biológicas e da Saúde* 17.1 (July 2011), pp. 37–45. DOI: `10.5212/Publ.Biologicas.v.17i1.0004`.

[238] Optimal Workshop. *Interpreting the pietree*. 2024. URL: `https://www.optimalworkshop.com/101guides/tree-testing-101#interpreting-the-pietree`.

[239] Nishigandha Deshmukh et al. 'Waste to nutrition: The evolution of whey, a byproduct to galactooligosaccharides production'. In: *Food Chemistry Advances* 4 (2024), p. 100642. DOI: `10.1016/j.focha.2024.100642`.

[240] Gustavo Luis de Paiva Anciens Ramos et al. 'Chapter 19 - Whey: generation, recovery, and use of a relevant by-product'. In: *Valorization of Agri-Food Wastes and By-Products*. Ed. by Rajeev Bhat. Academic Press, 2021, pp. 391–414. DOI: `10.1016/B978-0-12-824044-1.00030-1`.

[241] Silvana Magalhães Salgado, Nonete Barbosa Guerra and Artur Bibiano de Melo Filho. 'Polpa de fruta congelada: efeito do processamento sobre o conteúdo de fibra alimentar'. In: *Revista de Nutrição* 12.3 (Sept. 1999), pp. 303–308. DOI: `10.1590/S1415-52731999000300009`.

[242] Ricardo A. Falbo et al. 'Ontology patterns: clarifying concepts and terminology'. In: *Proceedings of the 4th International Conference on Ontology and Semantic Web Patterns*. Vol. 1188. WOP'13. Sydney, Australia: CEUR-WS.org, 2013, pp. 14–26. URL: `https://ceur-ws.org/Vol-1188/paper_11.pdf`.

[243] Tim Berners-Lee. *Linked Data*. 2006. URL: `https://www.w3.org/DesignIssues/LinkedData.html`.

[244] Pascal Hitzler. ‘A review of the semantic web field’. In: *Communications of the ACM* 64.2 (Jan. 2021), pp. 76–83. DOI: `10.1145/3397512`.

[245] Umutcan Simsek et al. ‘A Knowledge Graph Perspective on Knowledge Engineering’. en. In: *SN Computer Science* 4.1 (Oct. 2022), p. 16. DOI: `10.1007/s42979-022-01429-x`.

[246] Umutcan im ek et al. ‘Knowledge Graph Lifecycle: Building and Maintaining Knowledge Graphs’. In: *Proceedings of the 2nd International Workshop on Knowledge Graph Construction (KGCW 2021) co-located with the 18th Extended Semantic Web Conference (ESWC 2021)*. Ed. by David Chaves-Fraga et al. Vol. 2873. CEUR Workshop Proceedings. Online: CEUR-WS.org, 6th June 2021. URL: `https://ceur-ws.org/Vol-2873/paper12.pdf`.

[247] Leonel Molero Pereira. ‘Modelo de formação de preços de commodities agrícolas aplicado ao mercado de açúcar e álcool’. Doutorado em Administração. São Paulo: Universidade de São Paulo, May 2009. DOI: `10.11606/T.12.2009.tde-04062009-155921`.

[248] Maristela De Mello Martins. ‘Análise da eficiência técnica da pecuária de corte para regiões brasileiras selecionadas - uma análise de fronteira estocástica’. Mestrado em Economia Aplicada. Piracicaba: Universidade de São Paulo, Sept. 2019. DOI: `10.11606/D.11.2019.tde-04092019-164530`.

[249] Nayara Figueiredo. Gado para exportação à China registrou avanços na cotação. July 2024. URL: `https://globorural.globo.com/pecuaria/boi/noticia/2024/07/preco-do-boi-gordo-fecha-semana-em-alta.ghtml`.

[250] Reuters. *Reuters Commodities Market News*. Find the latest stock market news from every corner of the globe at Reuters.com, your online source for breaking international market and finance news. Sept. 2024. URL: `https://www.reuters.com/markets/commodities/`.

[251] Denise Cavallini Cyrillo. *Introdução à Inflação*. Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo, 2017.

[252] Erwin Johan Albert Folmer and Yvonne Verdonk. *Linked data in beeld: Platform Linked Data Nederland*. Platform Linked Data Nederland, 2014.

[253] Platform Linked Data Nederland. *Who we are*. Sept. 2024. URL: `https://www.pldn.nl/wiki/Wie_we_zijn`.

[254] Diego Bustamante and Hideaki Takeda. 'SPARQL Generation with Entity Pre-trained GPT for KG Question Answering'. In: *arXiv preprint* (2024). arXiv: `2402.00969`.

[255] Caio Viktor S. Avila et al. 'Experiments with text-to-SPARQL based on ChatGPT'. In: *2024 IEEE 18th International Conference on Semantic Computing (ICSC)*. Laguna Hills, CA, USA: IEEE, Feb. 2024, pp. 277–284. DOI: `10.1109/ICSC59802.2024.00050`.

[256] Md Rashad Al Hasan Rony et al. 'SGPT: A Generative Approach for SPARQL Query Generation From Natural Language Questions'. In: *IEEE Access* 10 (2022), pp. 70712–70723. DOI: `10.1109/ACCESS.2022.3188714`.

[257] Liubov Kovriguina et al. 'SPARQLGEN: One-Shot Prompt-based Approach for SPARQL Query Generation.' In: *Proceedings of the SEMANTICS 2023 EU: 19th International Conference on Semantic Systems*. CEUR Workshop Proceedings, 2023. URL: `https://ceur-ws.org/Vol-3526/paper-08.pdf`.

[258] Snusmumr Soft. *GPT "Semantic Web Senior"*. 2024. URL: `https://app.aiprm.com/gpts/g-XCrCvo3ih/semantic-web-senior`.

[259] Leonardo Candela, Dario Mangione and Gina Pavone. 'The FAIR Assessment Conundrum: Reflections on Tools and Metrics'. In: *Data Science Journal* 23 (May 2024), p. 33. DOI: `10.5334/dsj-2024-033`.

[260] Luiza Petrosyan et al. 'FAIR degree assessment in agriculture datasets using the F-UJI tool'. In: *Ecological Informatics* 76 (2023), p. 102126. DOI: `10.1016/j.ecoinf.2023.102126`.

[261] Robert Huber and Anusuriya Devaraju. 'F-UJI: An Automated Tool for the Assessment and Improvement of the FAIRness of Research Data'. In: *Proceedings of the EGU General Assembly 2021*. Mar. 2021. DOI: `10.5194/egusphere-egu21-15922`.

[262] Anusuriya Devaraju et al. *FAIRsFAIR Data Object Assessment Metrics*. Version 0.5. Apr. 2022. DOI: `10.5281/zenodo.6461229`.

[263] IPBES. *The assessment report of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services on pollinators, pollination and food production*. IPBES, Dec. 2016. DOI: `10.5281/ZENODO.3402857`.

[264] Alexandra-Maria Klein et al. 'Importance of pollinators in changing landscapes for world crops'. In: *Proceedings of the Royal Society B: Biological Sciences* 274.1608 (Feb. 2007), pp. 303–313. DOI: `10.1098/rspb.2006.3721`.

[265] Debora Drucker et al. 'Collectively Working towards Plant-Pollinator Interactions Data Interoperability and Reuse: Lessons Learned from the WorldFAIR Project'. In: *Biodiversity Information Science and Standards* 8 (Nov. 2024). DOI: `10.3897/biss.8.141109`.

[266] J. A. Salim et al. *Guidelines and Recommendations for Publishing Agricultural-related Pollinator Data, v1.0*. WorldFAIR Project, 2023. URL: `https://rebipp.github.io/worldfair-agrobio`.

[267] Matthew Jones et al. *Ecological Metadata Language version 2.2.0*. 2019. DOI: `10.5063/f11834t2`.

[268] Megan E. Frederickson. 'Mutualisms Are Not on the Verge of Breakdown'. In: *Trends in Ecology and Evolution* 32.10 (Oct. 2017), pp. 727–734. DOI: `10.1016/j.tree.2017.07.001`.

[269] José Augusto Salim. 'Unificando dados de interações bióticas: terminologia, análise de dados, padronização e proposta de um esquema de dados para interações planta-polinizador'. Doutorado em Sistemas Digitais. São Paulo: Universidade de São Paulo, Oct. 2023. DOI: `10.11606/T.3.2023.tde-03012024-110904`.

[270] Filipi Miranda Soares. 'Princípios para a criação de uma extensão de metadados sobre interações ecológicas na agrobiodiversidade para o padrão Darwin Core'. Master Thesis. Universidade Federal de Minas Gerais, Dec. 2019. URL: `https://repositorio.ufmg.br/handle/1843/33387`.

[271] GBIF.org. *Use Case: Biotic Interactions - Sottunga Island Melitaea cinxia Population Study*. 2023. URL: `https://discourse.gbif.org/t/use-case-biotic-interactions-sottunga-island-melitaea-cinxia-population-study/3312`.

[272] Quentin Groom et al. 'Improving Darwin Core for research and management of alien species'. In: *Biodiversity Information Science and Standards* 3 (2019), e38084. DOI: `10.3897/biss.3.38084`.

[273] Chris Mungall et al. *oborel/obo-relations: June Release*. July 2023. DOI: `10.5281/zenodo.8110608`.

[274] Luísa G. Carvalheiro et al. *Provenance of WorldFAIR pilot data from: VisitationData_Luisa_Carvalheiro*. Version 0.1. Zenodo, Jan. 2024. DOI: `10.5281/zenodo.10530109`.

[275] Nick Balfour et al. *The Database of Pollinator Interactions (DoPI)*. 2023. URL: `https://www.sussex.ac.uk/lifesci/ebe/dopi/`.

[276] L. G. Carvalheiro, E. R. M. Barbosa and J. Memmott. ‘Pollinator networks, alien species and the conservation of rare plants: Trinia glauca as a case study’. In: *Journal of Applied Ecology* 45.5 (Oct. 2008), pp. 1419–1427. DOI: `10.1111/j.1365-2664.2008.01518.x`.

[277] Jorrit H. Poelen. *Nomer Corpus of Taxonomic Resources*. Version 0.6. Zenodo, Sept. 2022. DOI: `10.5281/zenodo.7065661`.

[278] Nomer and Elton. *Review of WorldFAIR Agricultural Plant-Pollinator Data Pilot: Carvalheiro, LG; Barbosa, E.R.M. and Memmott, J. 2008. Pollinator networks, alien species and the conservation of rare plants: Trinia glauca as a case study*. Zenodo. [Data set]. 2024. DOI: `10.5281/zenodo.10647856`.

[279] Nomer and Elton. *Review of WorldFAIR Agricultural Plant-Pollinator Data Pilot: Kasina M., Kimani I., Mulwa S., and Muliaro W. A review of the status of web-based African Plant-Pollinator Interaction data*. Feb. 2024. DOI: `10.5281/ZENODO.10647892`.

[280] Simon Hodson. *WorldFAIR (D2.2) WorldFAIR’s Experience with FIPs (second set of FAIR Implementation Profiles for each case study)*. May 2024. DOI: `10.5281/ZENODO.11236094`.

Agriculture runs on data, but that data rarely speaks the same language. This dissertation introduces a practical, process-driven Semantic Interoperability Framework that aligns heterogeneous agricultural data using metadata schemas, ontologies, and knowledge graphs, with LLM-assisted workflows, to make FAIR by design a reality. Across two real-world case studies, the framework resolves inconsistencies, turns metadata into computable descriptions, and strengthens knowledge representation so that datasets can be discovered, linked, and reused. The result is a clear blueprint for harmonized data and interoperable information systems across the agri-food ecosystem. In short, it turns isolated datasets into a connected, machine-actionable knowledge base.

ISBN 978-90-365-6762-6

USP

UNIVERSITY
OF TWENTE.