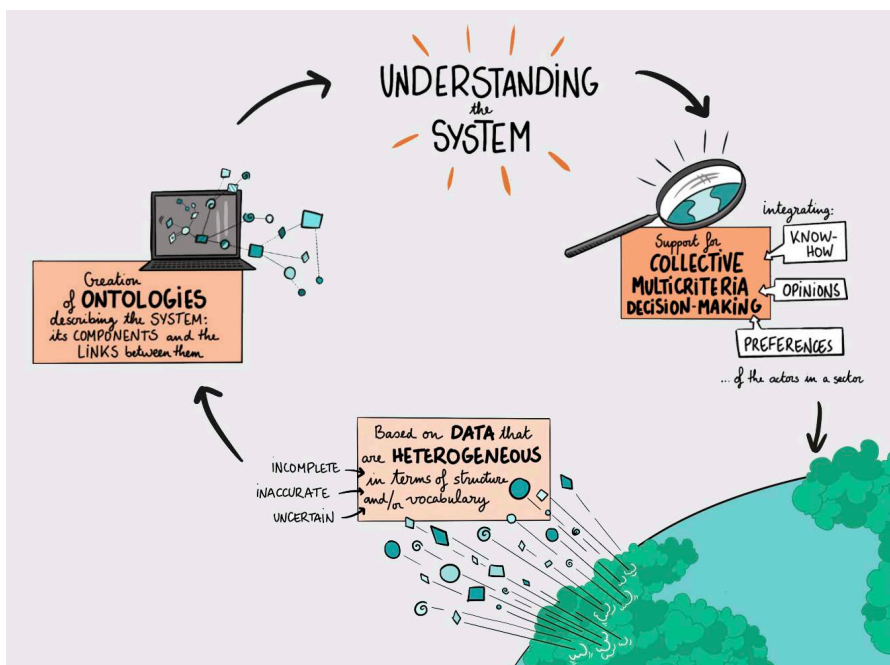


KNOWLEDGE ENGINEERING FOR DECISION SUPPORT IN THE FOOD AND BIOPRODUCT INDUSTRY

Integration of heterogeneous data
through AI techniques

Patrice Buche, Pierre Bisquert, Rallou Thomopoulos, eds



Knowledge Engineering for Decision Support in the Food and Bioproduct Industry

Integration of Heterogeneous Data
Through AI Techniques

Patrice Buche, Pierre Bisquert,
Rallou Thomopoulos, eds

Éditions Quæ

Quæ books in English

Remote Sensing and Spatial Modelling.

Applications to the Surveillance and Control of Mosquito-borne Diseases

Annelise Tran, Eric Daudé, Thibault Catry (Eds), 2025, 148 p.

<https://doi.org/10.35690/978-2-7592-4102-6>

Understanding Digital Agriculture.

Ten Years of Interdisciplinary Research at the #DigitAg Institute

Véronique Bellon-Maurel, Karine Gauche, Martha Lucia Enriquez, Nathalie Lyon-Caen (Eds),

2026, 202 p. <https://doi.org/10.35690/978-2-7592-4360-0>

To cite this work:

Buche P., Bisquert P., Thomopoulos R. (eds), 2026. *Knowledge Engineering for Decision Support in the Food and Bioproduct Industry. Integration of Heterogeneous Data Through AI Techniques*, Versailles, Éditions Quæ, 220 p. <https://doi.org/10.35690/978-2-7592-4317-4>

All book manuscripts are peer-reviewed by external reviewers.

This scientific evaluation process is described here:

<https://www.quae.com/store/page/199/processus-devaluation>.

The editorial process also relies on software for detecting similarities
and texts potentially generated by artificial intelligence.

Cover: Claire Hernandez, graphic facilitator.

© Éditions Quæ, 2026

Éditions Quæ

RD 10, 78026 Versailles Cedex, France

www.quae.com – www.quae-open.com

ISBN print: 978-2-7592-4316-7

ISBN ePub: 978-2-7592-4318-1

ISBN PDF: 978-2-7592-4317-4

ISSN: 1777-4624

The digital versions of this book are released under a CC-by-NC-ND 4.0 license
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>).



Contents

Foreword	5
<i>Johnny Beaugrand</i>	
Foreword	7
<i>Nathalie Aussenac-Gilles</i>	
Introduction	11
Chapter 1. Domain Ontologies for Food and Biobased Products	15
<i>Magalie Weber, Julien Cufi, Juliette Defossez, Stéphane Dervaux, Hervé Guillemin, Nicolas Guinet, Yannick Guyonvarch, Liliana Ibanescu, Emmanuele Piaud, Patrice Buche</i>	
1.1. Introduction	15
1.2 Proposed solutions	21
1.3 Outcomes.....	40
1.4 Conclusion and resources	47
Chapter 2. Task Ontologies for Product Processing and Characterization Processes	50
<i>Liliana Ibanescu, Julien Couteaux, Julien Cufi, Stéphane Dervaux, Hervé Guillemin, Alrick Oudot, Emmanuele Piaud, Magalie Weber, Patrice Buche</i>	
2.1 Introduction	50
2.2 Proposed solutions	55
2.3 Outcomes.....	85
2.4 Conclusion and resources	88
Chapter 3. Methods for Ontology-Guided Extraction of Experimental Data in Scientific Articles	92
<i>Martin Lentschat, Patrice Buche, Julien Cufi, Luc Menut, Mathieu Roche</i>	
3.1 Introduction	92
3.2 Proposed solutions	96
3.3 Outcomes.....	111
3.4 Conclusion and resources	117
Chapter 4. Reasoning About Process: Experimental Data and Technological Expert Know-how for Decision Support	121
<i>Mélanie Münch, Cédric Baudrit, Julien Cufi, Julien Couteaux, Christophe Fernandez, Luc Menut, Alrick Oudot, Patrice Buche</i>	
4.1 Introduction	121
4.2 Proposed solutions	126
4.3 Outcomes.....	135
4.4 Conclusion and resources	144

Chapter 5. Methods for Collective Decision Support that Consider both Preferences and Contradictory Arguments from Agrifood Chain Stakeholders	148
<i>Pierre Bisquert, Rallou Thomopoulos</i>	
5.1 Introduction	148
5.2 Proposed solutions	150
5.3 Outcomes	163
5.4 Conclusion and resources	168
Chapter 6. Designing Prospective Scenarios at the Scale of a Value Chain	171
<i>Romy-Lynn Chaib, Catherine Macombe, Rallou Thomopoulos, Erwan Engel</i>	
6.1 Introduction	171
6.2 Proposed solution	172
6.3 Outcomes	176
6.4 Conclusion and resources	179
Chapter 7. Assessing the Impact of Participatory Interventions at the Individual Level	181
<i>Reine Barbar, Rallou Thomopoulos</i>	
7.1 Introduction	181
7.2 Proposed solution	184
7.3 Outcomes	185
7.4 Conclusion and resources	185
Final Conclusion and Future Outlook	189
Appendix 1. Knowledge Engineering Models Used in the Book	192
<i>Jérôme Fortin, Patrice Buche, Pierre Bisquert</i>	
Appendix 2. Legend for the Workflow Diagrams	200
<i>Patrice Buche</i>	
References	202
Acknowledgments	216
List of authors	219

Foreword

In a world where data has become the lingua franca of science, business, and people, the ability to understand, link, and interpret information seems more critical than ever. Between the abundant quantitative data provided by sensors, experiments, and models from our labs and the wealth of qualitative knowledge produced by human expertise lies an imperfect interface that research is attempting to fix. It is exactly here that *Knowledge Engineering for Decision Support in the Food and Bioproduct Industry* comes in, striking a balance between rigorous measurements and nuanced meaning.

Based on an ambitious collective undertaking under the editorial coordination of Patrice Buche, Pierre Bisquert, and Rallou Thomopoulos, researchers at IATE (INRAE–Institut Agro–University of Montpellier), this book is first and foremost the story of a scientific and human adventure. It explores the many paths along which artificial intelligence, well beyond algorithmic performance, can become not only a tool for mediation between disciplines but also undoubtedly a lever of collective intelligence and hopefully a driver of responsible innovation in the agrifood and biobased sectors. This book offers a wide-ranging menu of options.

An integrative and open approach to knowledge

The first two chapters lay down the conceptual foundations of this multiauthor book. It presents the vision of open data integration, capable of linking experimental measurements, field observations, and expert knowledge. The authors show how semantic structures and ontologies – subjects on which they have been pioneers – provide the underlying architecture of meaning that guarantees interoperability and what could be called “the sustainability of knowledge.”

Chapters 3 and 4 get to the heart of the methodological approach. They describe the techniques of artificial intelligence that are used, such as machine learning, symbolic reasoning, knowledge representations and natural language processing, and above all the way these tools are combined in a hybrid, explainable, and participatory approach, which enhances their potential. Readers will discover the finesse of data engineering that never separates the technique from the human. The fourth chapter in particular illustrates the power of the methodological solutions put forward through a series of tangible applications in food and biobased products. From characterizing biomass to formulating new materials, each case study demonstrates how combining the quantitative and qualitative creates new opportunities for decision-making, traceability, and innovation.

Chapters 5 and 6 explore a common shortcoming faced by many labs and researchers – the organizational and collaborative aspects of a project – and how bridges can be created between labs, companies, and policymakers. These chapters also look at how open platforms can be built where data, knowledge, and preferences can become a common good rather than simply a technical instrument. Reading between the lines,

we can discern the authors' conviction that to be successful, a scientific project depends as much on the quality of the algorithms as it does on teamwork.

Finally, the last chapter expands the conversation to the future. It examines the role of open and smart systems in the digital transformation of the bioeconomy sectors and the ethical responsibility of their designers. More than a conclusion, it is an invitation: an invitation to continue the dialogue between fields and stakeholders, to create tools that are ever more transparent, collaborative, and meaningful.

A book that encourages dialogue between science and society

This book goes much further than simply presenting results. It embodies a way of doing science differently, openly, thoughtfully, and firmly rooted in the reality of current transitions and the need for research that seeks solutions to the climate emergency we now find ourselves in. This book is for researchers, engineers, and students, but also business leaders and public policymakers trying to understand how artificial intelligence can drive the sustainable transformation of food and biobased systems.

Johnny Beaugrand
TRANSFORM Division Head
INRAE

Foreword

At a time when large neural network models seem to be the solution to all problems and are often considered the only effective contributions of artificial intelligence (AI), I am delighted and honored to preface this book, which perfectly illustrates the richness, diversity, and relevance of AI research far beyond the use or development of these models alone. In the approach put forward through the work of the ICO team at the IATE–INRAE–Institut Agro–University of Montpellier unit, human knowledge and practices are sources of inspiration, objects of transmission, and key elements in the completion of tasks in conjunction with digital resources, whether software, data, or knowledge graphs. The work reported in this book is an exemplary contribution to knowledge engineering, both because it provides insight into a team research process and because it presents remarkable and reusable issues and results. It also contributes to the agrifood domain with a large set of real case studies illustrating how decision-making can be supported for tasks like food processing or food packaging design.

The editors of this book, Patrice Buche, Pierre Bisquert, and Rallou Thomopoulos, are renowned researchers in the field of AI, particularly in knowledge representation and engineering and know-how modeling, mainly in the agrifood sector. Patrice Buche has spent his entire career in this field, first as a lecturer at AgroParisTech until 2002, and since then as a research engineer at the IATE unit. He specializes in ontology alignment and knowledge extraction from texts and tables with a view to constructing ontologies and knowledge graphs used in decision support systems. Pierre Bisquert joined the IATE unit as an INRAE researcher in 2014, bringing his skills in reasoning, logic, and argumentation, mainly existential logic, to develop approaches based on knowledge graphs and rules for data integration. He took over from Buche as head of the ICO team. Finally, Rallou Thomopoulos has been conducting AI research in the agrifood sector since obtaining her PhD from AgroParisTech. As researcher and then research director at INRAE within the IATE unit since 2004, she has focused on the representation and interrogation of fuzzy or imprecise knowledge, then on the representation of chains of arguments in the form of graphs within decision support systems.

The work presented in this book is the result of the efforts of the ICO team at the IATE unit, of which all three editors are members. It contributes to the team's two areas of research: first, the ontology-driven integration of heterogeneous data and knowledge to formalize the knowledge needed for decision-making, and second, argumentation and social choice to integrate the preferences of the actors concerned and manage their different points of view. I should emphasize that it is rare for a research group to be able to capitalize so coherently on the work of each of its members. This book is the result of a long-term and genuine team effort, with each chapter itself a collective achievement. It confirms the complementarity and consistency of the work carried

out over the years, namely the projects and theses conducted within the ICO team. The permanent members of the team, and especially the editors of this book, played a key role in orchestrating, giving meaning to, and choosing a relevant direction for these different contributions, and now in bringing them together in this work. We can also be proud that this English edition reflects the original organizational structure of the French research environment, where a team includes several researchers working on personal scientific projects in conjunction with a team goal. In a way, this book could even be considered an ambassador for this approach. Indeed, it shows us that this way of working, which differs from the more pyramidal and individualistic organization of the English-speaking world, creates specific dynamics, promotes stimulating exchanges, and leads to outstanding results.

The seven chapters that make up this book provide a coherent and incremental reading experience. The authors are current and former members of the ICO team and long-term collaborators. Each chapter describes a building block that illustrates how AI can support decision-making in the agrifood sector, from collecting data and factual knowledge to building models, creating support tools, and evaluating the impact of proposed solutions on individuals, teams, and value chains. The results of one chapter are later reused in more ambitious contributions in subsequent chapters. Reading is made easier by the consistent structure of each chapter, which details the research issues and proposed solutions before presenting the outcomes that can be reused by the scientific community.

Chapter 1 (Domain Ontologies for Food and Biobased Products) presents various research projects that integrate multidimensional data into knowledge graphs and align the vocabularies used in different heterogeneous data sources. These solutions rely on domain ontologies. The importance of the terminological aspect of these ontologies (labels, comments, and definitions) for better validation by domain experts and more efficient alignment with similar resources or domain documents is worth noting. The outcomes of this work are not only the ontologies themselves, but also the methods and tools used to manage them (e.g., the PO² vocabulary and data manager) and to explore multiple databases (MultiDB explorer), classify data, and align source content.

Chapter 2 (Task Ontologies for Product Processing and Characterization Processes) introduces task ontologies, which are ontologies dedicated to specific tasks. These ontologies focus on the processes studied by the ICO team, such as product processing (e.g., bread or flour production from wheat) and product quality assurance and maintenance. Again, the outcomes include the ontologies, processes, and tools used to design them, as well as how these ontologies are populated to represent both the processing processes described in documents and data tables and the operational know-how that is typically passed on orally in organizations.

Chapter 3 (Methods for Ontology-Guided Extraction of Experimental Data in Scientific Articles) describes methods for the ontology-guided extraction of experimental data from scientific articles, with this work illustrated using scientific papers about food packaging. These methods rely on natural language processing to extract semantic relations. The work makes an original contribution to the extraction of n-ary relations (relations with more than two arguments), which can be used to collect several properties of a target entity. The process involves two steps: identifying the arguments and combining them to represent an instance of an n-ary relation.

Chapter 4 (Reasoning About Process: Experimental Data and Technological Expert Know-how for Decision Support) takes things a step further. Given the ontologies, knowledge graphs, and related methods and tools, the objective is to compile information on how this knowledge is utilized in the reasoning process for decision-making purposes. The three proposed approaches rely on both data and human expertise. They use either tree-based decision-making models, models trained on databases, or flexible queries based on fuzzy sets to combine several expert models. This work is applied to recommendations to correct food defects, select food packaging, and support biocomposite packaging design processes.

In Chapter 5 (Methods for Collective Decision Support that Consider both Preferences and Contradictory Arguments from Agrifood Chain Stakeholders), the objective targets the decision-making process itself. The proposed solution is based on the representation of the process using arguments in accordance with argumentation theory and social choice, and the decision-making process itself is examined. Two tools are presented: DAMN (a methodology for representing formal arguments in a statement graph and computing a decision by traversing this graph) and MyChoice (a web-based application that supports collective decision-making by assessing the acceptability of different alternatives based on several explanatory arguments for these alternatives).

Chapter 6 (Designing Prospective Scenarios at the Scale of a Value Chain) broadens the scope even further, evaluating innovations (i.e., changes in the value chain) by assessing their value to society. Three types of scenario-based methods are proposed to deal with standard situations, adapt to remote contexts, and optimize the evaluation process. The approach considers the technical feasibility and acceptability to humans of the proposed decisions. One example of a project involves improving food safety for children, with an evaluation of the overall impact of the new food chain.

Finally, Chapter 7 (Assessing the Impact of Participatory Interventions at the Individual Level) shifts the focus from the global impact to the local impact on individuals, using personalized food as an example. This assessment is based on a workshop in which visually impaired participants are informed and encouraged to prepare their own food. This has a significant impact on their knowledge, self-efficacy, practical skills, and cooking habits.

In conclusion, I highly recommend reading this book for several reasons, both for the scientific quality of its original contributions and for the rigor of the scientific approach it presents. It demonstrates the need to combine symbolic and more automated learning-based approaches in AI to achieve scalable systems that truly serve their users and provide relevant assistance. It will also be of interest to readers in the agrifood sector, who will find avenues for further work and significant contributions to the transfer of know-how and the advancement of knowledge in this field.

Nathalie Aussenac-Gilles
CNRS Research Director
Head of the MELODI team at IRIT

Introduction

This book explores real examples of applications to propose “workflows,” i.e., combinations of data and knowledge processing, to solve complex problems specific to the food and bioproduct industry. These workflows make it possible to collect, organize, share, connect, and regularly update different types of data and knowledge (e.g., expert opinions, experimental data, stakeholder preferences and opinions, digital models). The book tackles the methodological obstacle of reconciling knowledge and often imperfect data (of variable reliability, imprecise, incomplete or contradictory, and from heterogeneous sources) with learning, prediction, and decision-support methods and tools. Knowledge engineering, the main family of approaches used in this book, is the branch of artificial intelligence that deals with the acquisition, formal representation, and operationalization of expert knowledge for use in computer systems. It provides the methodologies and tools necessary to build knowledge-based systems by structuring domain expertise into formal models that support automated reasoning and decision-making (Pagels et al., 1984; Studer et al., 1998). This novel compilation of knowledge engineering methods and tools for the agrifood industry that allow semi-automatic reasoning on both quantitative data and qualitative knowledge constitutes the book’s originality. It results from a selection of research conducted by the INRAE knowledge engineering team (known by its French acronym ICO) at the IATE Research Unit in Montpellier, France, and its network of collaborators, namely within CATI DIISCICO, a professional IT network that brings together engineers and scientists from INRAE. This fruitful collaboration led the team and its network to win the Open Science Research Data Award in 2023 for the category “Creating the right conditions for reuse” from the French Ministry of Research. The research presented in this book by computer scientists specialized in knowledge engineering is the result of a long-term collaboration with many agrosresource processing experts belonging to the INRAE TRANSFORM department and associated partners. The authors of this book would like to take this opportunity to thank them for their support in helping us design the solutions proposed here. This research was also inspired by a long-term collaboration with cheesemaking value chains, and especially French Protected Designation of Origin (PDO) value chains. This collaboration led to the creation of the DOCAMEX association in 2023 at the annual assembly of the CNAOL, a national association for producers of geographically protected cheeses. Today, it brings together 20 PDO cheese value chains (70% of national production) and several ENIL dairy schools to ensure the maintenance of the co-produced DOCAMEX software tool by the ICO team with collaborators within CATI DIISCICO.

To cope with rapid changes in the agrifood sectors and the multifaceted challenges they face (environmental, societal, economic), stakeholders and decision-makers need a comprehensive and integrated view of the studied fields of application – the agrifood and bioproduct sectors – while being able to recommend actions at fairly detailed levels (e.g., optimal controlled parameter settings for biomass processing unit operations to achieve a given set of objectives, selection of innovative packaging

materials for a given food product, recommendation of technological actions to obtain a desired quality or to correct a defect). Decision-making must be based on available knowledge (data, expert statements, preferences and opinions of actors, numerical models) using a multicriteria approach, and it must consider the different and potentially conflicting points of view and preferences of the involved actors. This book presents a synthetic overview of workflows that have several specificities: (i) they are based on methods derived from artificial intelligence, (ii) they are implemented in innovative tools for capitalization and processing of data and knowledge, (iii) they are used to solve complex problems specific to the food and bioproduct industry, and (iv) their ultimate aim is to provide relevant collective decision-making support tools. These workflows are illustrated in several complex decision support applications (design of biocomposites for food packaging applications, food packaging selection for a given food, technological recommendations to preserve the quality of protected designation of origin cheeses, etc.).

This book focuses on the methods and tools used to implement these workflows. This presentation on two levels not only highlights the genericity of the proposed approaches but also ensures the work will remain relevant over time since it will not be dependent on a particular computer technology. Open resources are also provided (software sources, web/desktop applications, datasets, ontologies, thesauri, etc.) so readers can put the concepts presented in the book into practice and reuse/adapt them for other uses in the food and bioproduct industry or for other applications.

All the chapters are presented in the same way and broken down into four sections. Each chapter begins with a section that presents issues and challenges and identifies the research questions addressed in the chapter. Workflow solutions proposed to tackle those questions are compared to the state of the art. In Section 2, workflow solutions are detailed for each research question. Outcomes associated with the solutions are presented in Section 3. Section 4 concludes the chapter and lists associated resources available on a public repository (e.g., Recherche Data Gouv, INRAE GitLab, HAL, INRAE Nextcloud). For all chapters, 34 main scientific papers, 7 domain ontologies, 3 task ontologies, 12 data papers comprising 43 datasets, 10 other datasets, 16 software programs, 7 web/desktop applications, 15 videos, and 11 reports are available through open access.

The first part of this book, which is divided into four chapters, covers the integration of heterogeneous data sources guided by ontology and coupling with multicriteria decision support systems:

1. Domain Ontologies for Food and Biobased Products
2. Task Ontologies for Product Processing and Characterization Processes
3. Methods for Ontology-Guided Extraction of Experimental Data in Scientific Articles
4. Reasoning About Process: Experimental Data and Technological Expert Know-how for Decision Support

The second part, with three chapters, is dedicated to collective decision support:

5. Methods for Collective Decision Support that Consider both Preferences and Contradictory Arguments from Agrifood Chain Stakeholders
6. Designing Prospective Scenarios at the Scale of a Value Chain
7. Assessing the Impact of Participatory Interventions at the Individual Level

Interested readers will find a short introduction to logical foundations and standardized languages used in the chapters to implement proposed solutions in Appendix 1. The graphical representation of workflow solutions used in all chapters is presented in Appendix 2.

Chapter 1

Domain Ontologies for Food and Biobased Products

Magalie Weber, Julien Cufi, Juliette Defossez, Stéphane Dervaux, Hervé Guillemin, Nicolas Guinet, Yannick Guyonvarch, Liliana Ibanescu, Emmanuele Piaud, Patrice Buche

► 1.1. Introduction

1.1.1. Issues and challenges

Knowledge discovery in food and biobased products engineering is challenging because the research questions that must be addressed focus on both process performances and product uses. Examples of complex research questions include food or biobased product formulation, process optimization, and the environmental impact assessment of processes. Around the world, momentum is also building to bring production and consumption patterns together in a circular, more sustainable approach, known as the biobased economy or bioeconomy. According to the Organisation for Economic Co-operation and Development (OECD, 2021), the agrifood system is actually a multidimensional system that encompasses “all the elements and activities related to producing and consuming food, as well as their effects, including economic, health, and environmental outcomes.” Biorefining is defined by the IEA Bioenergy Task 42 on Biorefineries (de Jong et al., 2012) as “the sustainable processing of biomass into a spectrum of marketable products (food, feed, materials, chemicals) and energy (fuels, power, heat).” The need to develop more integrative approaches to reconnect agriculture, environment, food, and health has been reaffirmed recently to cope with the challenges of agrifood system sustainability and the “OneHealth” approach.¹

The science at play in connecting these different areas involves various disciplines, from agronomy to environmental sciences, socioeconomics, biochemistry, microbiology, nutrition, sensometrics, chemistry and physics, process engineering, and computing sciences. The data collected are by nature multidimensional, and large collections of experimental datasets, which come from various projects and with heterogeneous sources and formats, should be exploited. These data are also sometimes incomplete or imprecise due to the high uncertainty of measurements inherent in biological systems or analytical techniques. As a result, these data must be integrated, which involves giving them a unified structure and unified semantics before

1. <https://www.inrae.fr/en/europe-and-world/one-health-people-animals-and-environment>

modeling and reasoning. Moreover, additional external data are needed to enrich the datasets. This is the underlying motivation for the use of data matching and alignment techniques to identify vocabularies with similar descriptions in heterogeneous data sources. In this chapter, we look at two real examples to illustrate this objective. In the first example, the goal is to complete missing data (vitamin C, vitamin B12, and iron in food) from one source (CIQUAL, the French food composition database) using an external source (USDA FoodData Central database). In the second example, the aim is to track changes in food composition (e.g., sugar and salt content) found in heterogeneous sources by drawing from a common reference managed by OQALI (the French food observatory). The results obtained using these methods are given in Sections 1.3.2 and 1.3.3. In this chapter, classification and alignment are considered to be the same task in which two types of alignments are distinguished: equivalence or specialization.

We address two research questions in this chapter:

- Q1.1: How can the multiple dimensions of the agrifood system be integrated to design sustainable, healthy, and affordable food and biobased products?
- Q1.2: How can vocabularies defined in different data sources be matched for integration purposes?

1.1.2. State of the art and positioning, including presentation of available scientific papers describing our proposal

Ontologies are a powerful tool for integrating the many dimensions of food systems. An ontology is a formal representation of a set of concepts within a domain and the relationships between those concepts (Gruber, 1993; Uschold and Gruninger, 1996). Ontologies are used to model knowledge in a way that allows mutual understanding between experts and interoperability among knowledge-based systems and agents (not only humans but also machines). Ontologies enable the standardization of vocabulary and provide a framework for data integration, reasoning, and analysis. There is a broad spectrum of ontologies, ranging from top-level ontologies (foundational ontologies), intermediate-level ontologies defining basic models or generic standards (core ontologies or ontology design patterns), and application-level ontologies (task or domain ontologies). Top-level ontologies are fundamental models, which are the most formal abstractions. A core ontology defines the generic model that can be further specialized for different domains or fields of applications. A task ontology focuses on the actions or tasks that should be performed to achieve a goal, often independent from a specific domain. A domain ontology gathers concepts from a specific domain.

In this book, we focus on and distinguish different kinds of ontologies, namely core ontologies (intermediate level) and domain or task ontologies (application level). For example, the core ontology for product processing might comprise both concepts (e.g., transformation steps and components) and relationships (e.g., *hasInput* (resp. *hasOutput*)) that link an input (resp. output) component to a step in a generic way. A domain ontology focuses on the entities and relationships specific to a particular area of knowledge (the domain of interest). For instance, a food domain ontology might include concepts such as raw materials (e.g., wheat, milk), final products (cheesecake), nutrients (e.g., vitamin C), and unit processing operations (e.g., heating), and relationships between entities (e.g., durum wheat is a kind of wheat, or cheesecake is a kind of cake).

The core ontology and the task ontology can be intertwined, as the generic concepts of the core ontology can both define tasks and be specialized within a domain ontology for a given application. For example, the core ontology presented above was designed to represent processing tasks. Tasks are expressed in a generic way in the core ontology, and concepts from the domain ontology are more specific to the domain (e.g., wheat is a kind of component, and heating is a kind of processing step).

Research question Q1.1 is addressed by using a domain ontology, TransformON, which is described in Section 1.2.1. This ontology will cover the aspects related to the processing of food, feed, bioproducts, and biowaste, as well as the characterization of the properties of raw materials, primary derivatives, end products, and waste or residue from the treated biomass.

To build the domain ontology, we used the Linked Open Terms methodology (Poveda-Villalón et al., 2022). First, we considered the existing resources needed to cover the defined domain, and more specifically both food and nonfood applications of biomass processing, including food, feed, biobased products, and biowaste valorization into bioactive molecules or biobased compounds, matter, and energy.

Following a literature study, we found many existing classification and description systems about food (see Ireland and Møller, 2016, for a review). Food classification and food description systems have different goals. A classification system tends to group foods from broad to narrower terms with similar characteristics. “Milk,” for example, can be classified under “dairy products,” while “wine” would go under “alcoholic beverages.” This system serves as a solution for facilitating information retrieval for end users. Meanwhile, a description system is a solution for the data originator/producer, who wants to describe the food as precisely as possible from different points of view, without the need to aggregate them. For example, the International Network of Food Data Systems (INFOODS) (Truswell et al., 1991) provides free-text food descriptions, a listing of features or entities that might influence food composition and that data providers are encouraged to record. However, it does not always include controlled terms to standardize the description and facilitate data retrieval. In this type of situation, thesauri can be used to ensure consistent indexing and retrieval (ISO, 2011, 2013). A thesaurus, in which vocabulary control is achieved by providing textual definitions and organizing hierarchical relationships, is a useful resource that can complement a coding system.

Here are some valuable existing resources for the food domain:

- *The AGROVOC² and the LanguaL™ thesauri* are well-known thesauri for the agrifood domain. AGROVOC is a controlled vocabulary covering all areas of interest to the Food and Agriculture Organization of the United Nations (FAO), including food, nutrition, agriculture, fisheries, forestry, and environment. AGROVOC was first developed in the early 1980s, and currently contains over 41,000 concepts organized in a faceted hierarchy; each concept may have labels in up to 42 languages. However, its facets and terms are not sufficient or specific enough to describe food products in the form in which they are consumed. LanguaL™ (Ireland and Møller, 2010) is an international thesaurus using faceted classification and whose origins stretch back to the 1975 US FDA CFSAN Factored Food Vocabulary. This vocabulary, known by its acronym FFV, was first used in an information retrieval system

2. <https://www.fao.org/agrovoc/releases>

called “SIREN,” the Scientific Information and Retrieval Exchange Network, at the Center for Food Safety and Applied Nutrition (CFSAN) of the US Food and Drug Administration (FDA). The current version of LanguaL™ 2017 contains a total of 12,605 descriptors arranged in a hierarchical structure organized into 14 facets. In all, more than 40,000 foods from Europe, North America, and a number of countries are indexed in LanguaL™. For example, the USDA National Nutrient Database for Standard Reference (SR legacy) has been fully LanguaL™ indexed, whereas the CIQUAL French Food Composition Database³ published by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) has included only the facet A to reference the product category for all foods and has only fully LanguaL™-indexed a subset of foods.

– *FoodEx2*, which was developed by the European Food Safety Authority (EFSA), is “a comprehensive food classification and description system aimed at covering the need to describe food in data collections across different food safety domains” (EFSA, 2015). The general aim was to provide the food details necessary to make exposure or intake assessments in a food safety context. FoodEx2 is an authoritative reference that complements the EFSA’s Standard Sample Description for Food and Feed (EFSA, 2013) at the European level and is used by national agencies in European countries for database annotation and data exchange. A parent-child relationship exists between a core list food item and its related extended list food items. The terms in the core and extended lists may be aggregated in different ways according to the needs of the different food safety domains. Collections of additional terms (called “facet descriptors”) are also available. The list of terms describes the group to which the food product belongs, as well as the product itself. A unique term code is linked to each list term or facet descriptor, so it can be described in an unambiguous way. The classification is mostly based on the treatment used to obtain new food product categories and considers three types of food categories: raw materials, derivatives/ingredients, and composite foods. Raw materials are directly obtained from living sources after harvesting or slaughtering. Derivatives/ingredients are obtained by applying a process to raw materials. Composite foods are foods made with several raw materials or derivatives/ingredients. FoodEx2 also provides a list of feed for different groups of animals, which was also in our scope.

– *FoodOn*⁴ is an ontology built to be interoperable with the international Open Biological and Biomedical Ontology Foundry (OBO Foundry⁵) and to represent entities that play a “food role” (Dooley et al., 2018). When it launched in 2018, FoodOn was largely based on LanguaL™ categories that were further converted in the ontology environment. The process of creating logical definitions started with a large number of food products that were introduced by way of the LanguaL™-indexed SIREN database. FoodOn provides a controlled vocabulary that can be used by both people and computers to name all parts of animals, plants, and fungi able to bear a food role for humans and domesticated animals, as well as derived food products and the processes used to make them. The aim is to develop semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, and culinary, nutritional, and chemical ingredients and processes.

3. <https://ciqual.anses.fr/>

4. <https://FoodOn.org/>

5. <https://obofoundry.org/>

Other ontologies were also examined to cover our targeted scope. We selected two ontologies, CheBi⁶ for biochemical and chemical substances, and OBI⁷, the Ontology for Biomedical Investigations for the description of biochemical assays. Other OBO Foundry ontologies were also taken into account when appropriate terms were available. The NCBI taxonomy⁸ was also considered for living organisms. However, as we did not find ontologies to cover the whole scope, especially for nonfood bioproducts and biowaste, we also selected nonsemantic resources recognized as references such as the European Waste Catalogue (EWC), a hierarchical list of waste descriptions established by the European Commission Decision 2000/532/EC, and the IUPAC nomenclature⁹ for chemistry and physics.

Since a couple of domain ontologies were previously used in our research on dataset annotation, we also built the new ontology by re-engineering four existing domain ontologies (see Figure 1.1) in order to ensure backward compatibility with the existing datasets (Guillard et al., 2016; Lousteau-Cazalet et al., 2016; Fabre et al., 2020; Buche et al., 2021b; Munch et al., 2022a, 2023a). Data stewardship and the need for integration of other information relative to life cycle analysis of food and biobased products have been our driving forces.

To reconcile the different domain vocabularies, a definition of terms was drawn up with experts in each domain, using the nomenclatures and other authoritative references previously listed. This approach is part of the global Open Science/Open Data process, involving the implementation of FAIR (Findable, Accessible, Interoperable, Reusable) principles and semantic web technologies.

To answer the second research question Q1.2, we propose two approaches:

- When few alignments between heterogeneous data sources are available, we suggest aligning vocabularies using an unsupervised approach based on background knowledge.
- When there are many alignments between heterogeneous data sources available, we propose using a supervised machine learning approach.

Ontology alignment or matching is a process where entities (i.e., concepts or relations, which are object properties) of two ontologies are paired and characterized as being the same or similar with a given degree of confidence (Shvaiko and Euzenat, 2013). Experts in ontology alignment highlight the importance of using specific domain knowledge (Van Hage et al., 2010). Indeed, general external knowledge sources, such as the WordNet lexical database, lack the necessary semantic details to accurately identify matches between domain-specific concepts. Numerous studies have demonstrated the benefits of incorporating background knowledge in the ontology matching process (Aleksovski et al., 2006; Faria et al., 2013; Tigrine et al., 2016; Annane et al., 2018). Our method, called OntoMapF, presented in Buche et al. (2021a) and detailed in Section 1.2.2 in this chapter, can be seen as an ontology matching process that uses background knowledge encoded within the LanguaL™ thesaurus. It implements the unsupervised approach when few alignments between heterogeneous data sources are available. Ispirova et al. (2017) proposed an alignment method specifically designed for nutritional data in food sources.

6. <https://www.ebi.ac.uk/chebi/>

7. <https://pubmed.ncbi.nlm.nih.gov/27128319/>

8. <https://www.ncbi.nlm.nih.gov/taxonomy>

9. <https://iupac.org/>

They compared various string-matching techniques and concluded that the best results were achieved by combining part-of-speech (POS) tagging with probability theory. However, their method was limited to English food nutrient names, which are simpler to compare due to their standardized structure using chemical names, adjectives, and numbers. Food product names pose a greater challenge as they can include names that are specific to different countries or verbs describing processing methods, and they are generally more complex. The Ontology Alignment Evaluation Initiative (OAEI) included a “food” task in its 2006 and 2007 campaigns (Euzenat et al., 2006; Van Hage et al., 2010). This OAEI task drew from the AGROVOC and National Agricultural Library Thesaurus Concept Space (NALT) agricultural thesauri as external references.

The challenges identified for the OAEI food task were to:

- Align a large dataset
- Align concepts from diverse domains beyond just food
- Perform alignment based on the limited semantics of the thesaurus structure, requiring term disambiguation and effective lexical matching strategies

When the 2006 OAEI food task is compared with OntoMapF, our proposed method, some key differences are revealed:

- Our datasets are smaller
- Our concepts belong exclusively to the food domain
- We use the multifaceted LanguaL™ thesaurus, which requires a methodological approach that incorporates this information and which was not addressed in the OAEI food task

Moreover, we built two gold standards: one for the CIQUAL/FoodOn gold standard alignment¹⁰ and one for the CIQUAL/USDA gold standard alignment¹¹ to evaluate the performance of our data-linking algorithms and calibrate them for higher accuracy. These sets of gold standard data have been manually verified by two independent raters to avoid any bias. We then proposed a new food track for the OAEI food track campaigns for 2022, 2023, and 2024 (Pour et al., 2022, 2023, 2024) to compare our results to those obtained using state-of-the-art alignment methods. In the three campaigns, none of the evaluated matchers found all reference matches correctly. LogMapLt stood out for its fast computing speed, while Matcha produced the best results for the application. The use of background knowledge available in the CIQUAL and SIREN/USDA ontologies in terms of food description based on FoodOn concepts should be considered in future OAEI campaigns. Our OntoMapF approach outperformed all of the algorithms used in the OAEI competition.

Our method Oqlassif¹² (Defossez et al., 2022), described in Section 1.2.2, can be used to categorize food products into classes defined in a predetermined thesaurus when many alignments between heterogeneous data sources are available. The thesaurus has two levels, referred to as sectors and families in the approach described below. Each sector has one or more families, and a single family cannot belong to more than one sector. Some papers for supervised food product classification have been proposed to solve similar kinds of questions (Eftimov et al., 2017; Hu et al., 2023). Eftimov et al. (2017) proposed a

10. <https://doi.org/10.15454/BVXD71>

11. <https://doi.org/10.15454/4XIBS9>

12. This method was created during a project financed by Oqali (<https://www.oqali.fr/>) and managed by the Odalim platform (<https://www.odalim.inrae.fr>) in close collaboration with the IATE and LIRMM units.

tool called StandFood to classify new candidate food products into four FoodEx2 classes representing different kinds of products (raw, derivative, simple composite, aggregated composite); the training dataset included 5,416 products. Hu et al. (2023) introduced a method able to categorize food product descriptions (name, trademark, ingredient) in 172 families of ingredients from Health Canada's table of reference amounts (Health Canada, 2024); the training dataset included 92,000 products. The aim of Oqlassif, an original contribution of this chapter, was to classify a given food product into 32 sectors and more than 600 families belonging to the OQALI database (Menard et al., 2011) thanks to an initial training set of 60,000 product descriptions composed of a set of textual descriptive variables. In new training, the method designed for Oqlassif must be able to integrate changes in terms of food products and OQALI categories to account for the French food market dynamic. Two original approaches were designed to do this. First, the selection of relevant descriptive variables characterizing the products to be given to our algorithm was fundamental, so OQALI experts were asked to select those variables. In comparison, the StandFood tool from Eftimov et al. (2017) needs less information as it classifies food products into only four immutable food categories (raw, derivative, simple composite, aggregated composite). Second, to classify a product description into a large number of classes (600+ families) in Oqlassif, a two-layer classification is implemented: first, descriptions are categorized into sectors, and then for each sector, a specific classifier predicts the family belonging to the sector. The most adequate classifiers (between random forest and support vector machines [SVMs] identified as the most relevant in preliminary tests) and their hyperparameters are selected every year using new food product entries from the previous year. Indeed, Oqlassif is used in production by the OQALI team. The classifiers are trained once a year, with new food product descriptions from the previous year being integrated. Decisions are made according to a 5-fold cross-validation algorithm upstream, which was possible because of our large training set. Eftimov et al. (2017) used several classifiers, some of which were also used in the Oqlassif method, and there is a majority vote between classifiers that has the final say. Hu et al. (2023) used the k-nearest neighbors approach with XGBoost and Manhattan distance. In terms of accuracy, the three approaches are similar: Eftimov et al. (2017) achieved 91.2% accuracy with the majority vote between classifiers, while Hu et al. (2023) achieved 98% for category classification and 96% for subcategory classification. Oqlassif achieved 93.4% accuracy with its two-layer classification. The Oqlassif method and associated tool are different from Hu et al. (2023) in the sense that they are able to take into account changes in the OQALI thesaurus during annual training. Oqlassif has a more sophisticated categorization system, with 600 subcategories grouped in 30 categories with classifiers optimized for each of those 2 category levels, compared to 4 categories for Eftimov et al. (2017) and 24 categories/172 subcategories for Hu et al. (2023).

►► 1.2 Proposed solutions

1.2.1 Solutions proposed to integrate the multiple dimensions of the agrifood system

To answer the first research question Q1.1, we relied on our expertise and previous developments in data integration in two RDF databases developed over the last 10 years through different research projects: the BaGaTel (Pénicaud et al., 2019) and @Web (Buche et al., 2013a) graph databases. These databases are based on two different task

ontologies: for BaGaTel, the Process and Observation Ontology (PO²) (see Chapter 2, Section 2.2.1), and for @Web, the naRyQ ontology (n-ary relations between quantitative experimental data) (see Chapter 2, Section 2.2.3). Those two task ontologies, software, and database implementations are discussed in more detail in Chapter 2.

Several domain ontologies specializing these two task ontologies were used for dataset annotation: (i) naRyQ milk microfiltration¹³, (ii) naRyQ biorefinery¹⁴, (iii) naRyQ transmat (matter transfer)¹⁵, and (iv) PO² Dairy gels.¹⁶

For the building/re-engineering of the domain ontology, we had to (i) transfer domain ontologies designed for naRyQ into the PO² model; (ii) take into account new competency questions for the PO² model; and (iii) build the new TransformON¹⁷ domain ontology based on the PO² core model.

In ontology engineering, competency questions (CQs) (Uschold and Gruninger, 1996) are natural-language questions that outline the scope of knowledge represented by an ontology and the applications exploiting it. CQs represent functional requirements in the sense that the ontology and the developed ontology-based information system should be able to answer them.

The main competency questions (CQs) identified when building PO² were:

- CQ1: Which steps compose a given processing process?
- CQ2: Which attribute values are associated with controlled/observed parameters for each step?
- CQ3: Which are the observations where a specific attribute is observed for a given input (or output)?
- CQ4: What are the changes for an attribute value of an input during a given step?
- CQ5: What are all the steps and observations executed by a specific person or organization?

The new CQs identified during the specification phase were:

- how to model a comprehensive process for biomass processing by being able to distinguish between food and nonfood products (CQ1 bis)?
- how to distinguish between what comes from primary production, secondary processing, and waste (CQ2 bis)?
- how to represent the experimental observations throughout the process by being able to identify the object of interest in the observation (CQ3 bis)?
- how to represent the list of types of equipment for a given platform (CQ4 bis)?
- how to retrieve the replications of a process with respect to an experimental design (CQ5bis)?
- how to identify metadata allowing traceability and harvesting of the ontology and corresponding datasets once published online (CQ6)?

To meet these requirements, the PO² core model (Weber et al., 2023), described in Section 2.2.1, integrated new relationships (hasObjectOfInterest, hasReplicate); new SOSA concepts (sosa:Platform and sosa:System) (Janowicz et al., 2019); the light-

13. <https://doi.org/10.15454/5MQMKG>

14. <https://doi.org/10.15454/X2MOWO>

15. <https://doi.org/10.15454/NK24ID>

16. <https://doi.org/10.15454/THZ9I1>

17. <https://doi.org/10.57745/DWX7W6>

weight ontology for sensors, observations, samples, and actuators recommended by W3C; and some DCAT properties.

To build the domain ontology, the existing vocabularies were harmonized and the different domain ontologies were restructured to produce a unified vocabulary based on the PO² ontology, which allowed us to represent both food and nonfood processing processes and the characterization of food, feed, bioproducts, and biowaste. This vocabulary serves as a pivot to annotate (i) internal data produced in the user's laboratory, and (ii) external resources from the literature.

The process of building TransformON is illustrated in Figure 1.1. The PO² ontology structure guides the whole process (Figure 1.1a). Indeed, the PO² ontology provides the main concepts and relations to represent a general processing process, composed of steps, with inputs and outputs and using materials and methods, each one being characterized by a set of observations (see Figure 1.2). First, in the top-down approach, the set of external candidate resources (selected nomenclatures and authoritative references selected to cover the domain, such as FoodEx2, the European Waste Catalogue, and IUPAC nomenclatures) was exploited to build the domain ontology. The domain experts selected the most appropriate terms and the ontology engineer proceeded in organizing them into a hierarchical structure conforming to the PO² model. Second, in the bottom-up approach, our own existing subdomain (or microdomain) ontologies (e.g., PO² Dairy gel, nAryQ milk microfiltration) were imported to ensure backward compatibility with existing datasets, and complementary terms were added to cover the scope of the new domain ontology.

The process of ontology enrichment and curation is detailed in the complementary part of the dotted boxed area in Figure 1.1b. Users of the PO² Manager Data annotation tool have the option to generate the list of candidate concepts needed to annotate their data and which are not available in the ontology. This list is transferred to a data steward, who checks the list of candidate concepts to propose terms and definitions for those concepts. The list of candidate terms and definitions are then proposed to the ontology custodian, who finally integrates them into the ontology using the PO² Manager Vocabulary tool maintained by the ontology developer.

PO² Manager (Munch et al., 2022c; Weber et al., 2023) is a Java application designed to assist domain experts on two tasks: (i) describing experimental processes using the concepts of the application domain thanks to the PO²Manager Dataset feature, and (ii) extending the PO² core concepts with concepts associated with the application domain using the PO²Manager Vocabulary feature.

The PO²Manager Dataset feature is a graphical user interface (GUI) designed to allow data entry using the TransformON domain ontology. The left side of Figure 1.2 shows the editor that enables the user to construct processes made up of unit operations as well as input and output components. For example, the "Itinerary 9" process in Figure 1.2 is composed of three sequential unit operations (coarse milling, drying, and fine milling). Hibiscus 1 is the input component of the coarse milling and Hibiscus Powder 1 is a powder that is the output component of the milling operation. The right side of Figure 1.2 shows the graphical representation of the process. The black circles represent sequential unit operations (steps) and the red circles represent input or output compositions. A black circle with green (9.3 Impact milling) means that observations are included in the step.

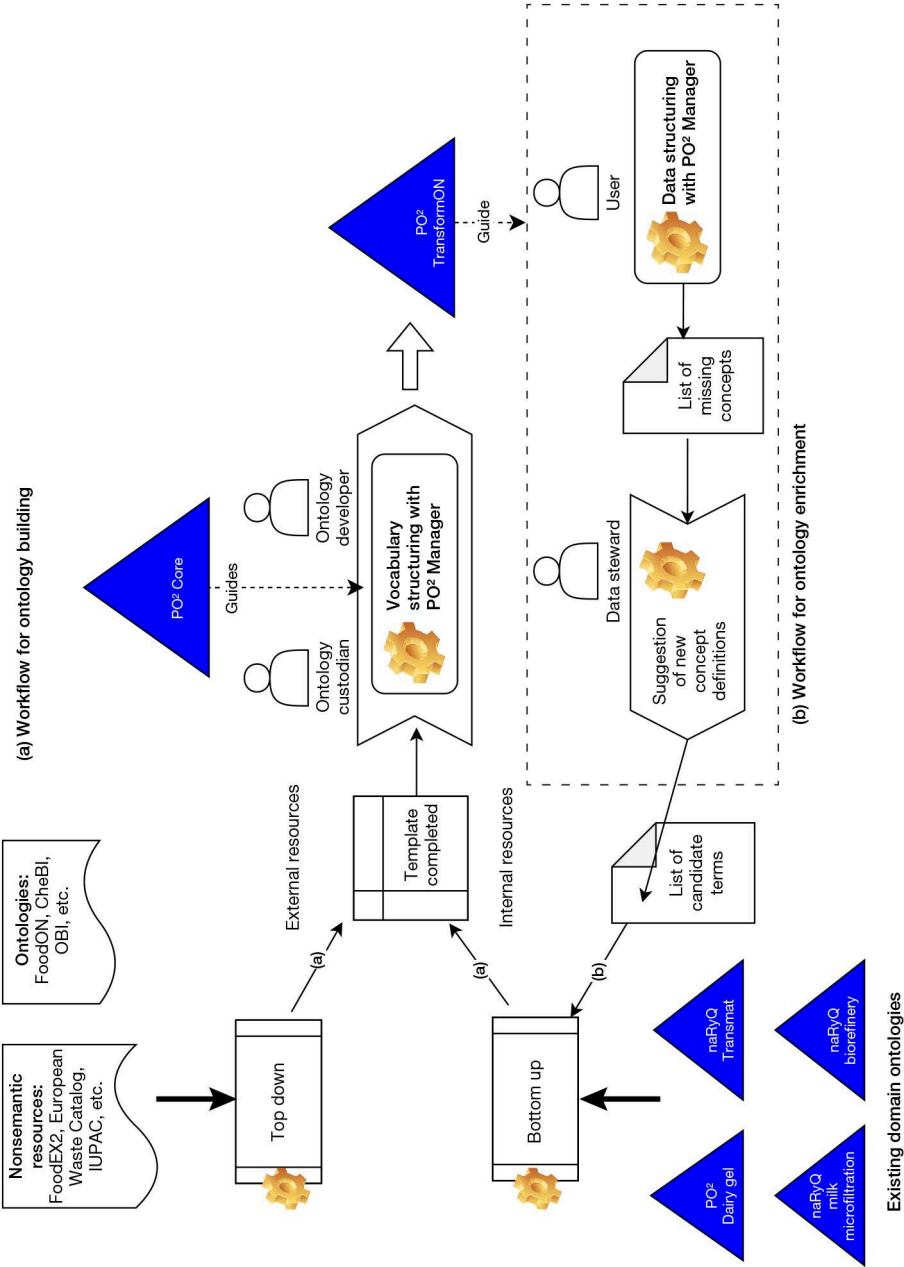


Figure 1.1. Workflows for TransformON domain ontology building and enrichment.

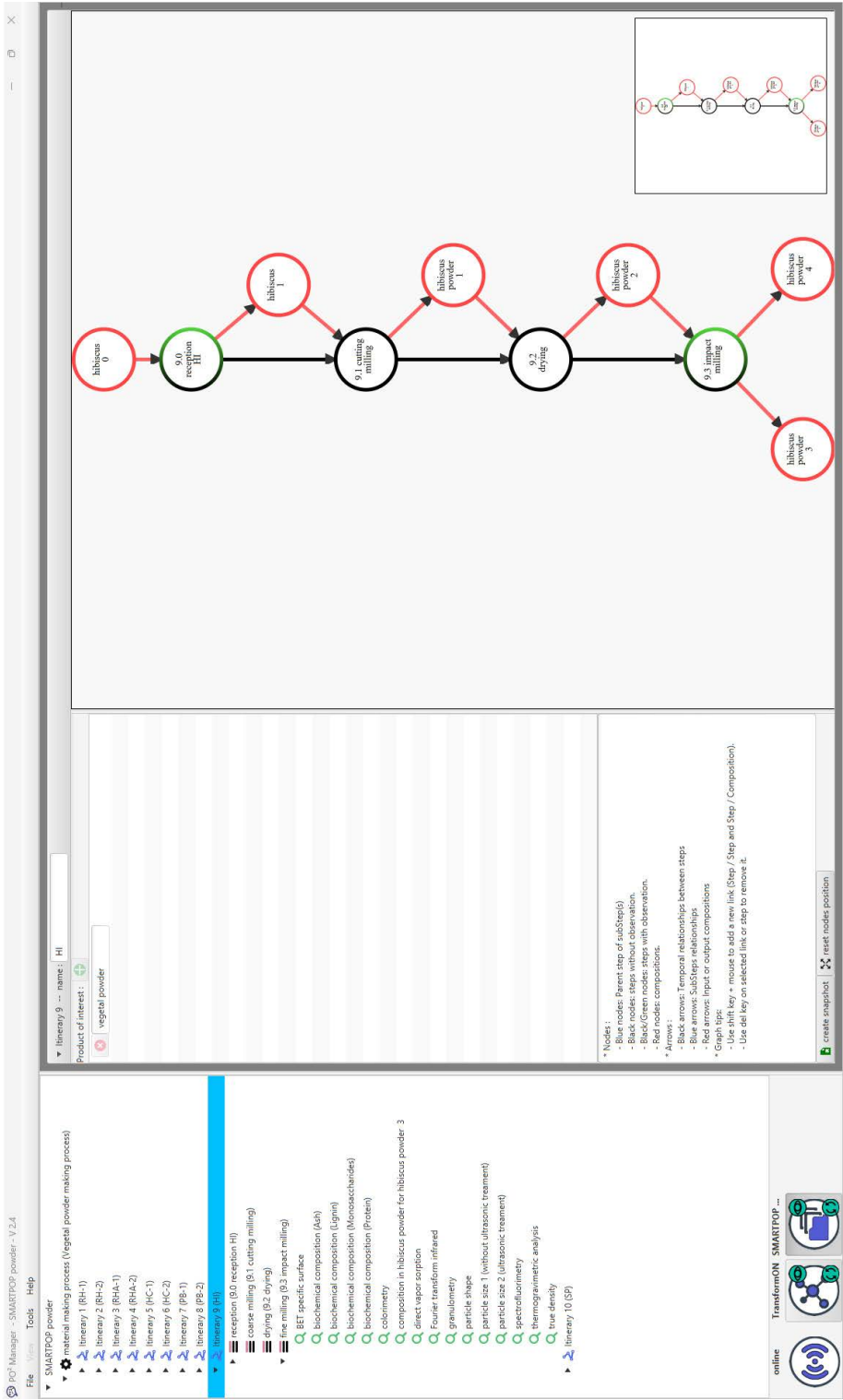


Figure 1.2. Screenshot showing the process editing with PO² Manager: a detailed look at the sequence of steps to transform the hisciscus plant into powder.

Observations include materials and methods as well as one or more data tables. Several observations and characteristics (e.g., biochemical composition, colorimetry, etc.) associated with the Hibiscus Powder 4 output component under the fine milling operation are shown on the left-hand side of Figure 1.3. On the right-hand side, the details of one observation are shown and include the metadata (dates, time, agents), scale, and materials and methods associated with the observation. Finally, the results of the observation are included in tables. In the example, there is just one table (Table 1), which contains one individual result (% ash rate).

Data represent instances that are linked to appropriate classes using the selected domain ontology (here, TransformON). Using PO² Manager, data structuring and annotation can be performed based on the PO² model and the controlled vocabulary provided in the TransformON domain ontology. Other metadata can also be recorded, such as the date, time, and duration (fields outlined in red in Figure 1.3), as well as the organization/agent performing the task. We will explain why data structuring and annotation are helpful in finding information in Chapter 2.

PO² Manager's Vocabulary feature also helps the ontology custodian edit the terminological part of the ontology and expand the PO² core concepts with concepts related to the application domain (see Figure 1.4). Each core concept in the PO² core ontology is the top concept of one hierarchy of classes. Each class of the hierarchy is also a Skos:Concept that allows us to manage different RDF properties: one preferred label per language and synonyms, definitions, and notes (ScopeNote, EditorialNote, or other Note). Concepts can be grouped into skos:ConceptScheme, representing a thesaurus in the SKOS format. Alignments can also be displayed (skos:ExactMatch or skos:CloseMatch). The ontology versioning is also guaranteed by the tool.

The window displaying the concept also shows its clickable URI ("http" URI or URL), which links to an HTML page thanks to the dereferencing mechanism. Dereferencing a URI means to request a representation of the resource designated by the URI in different formats according to the content negotiation. HTTP is an example of a protocol that enables representation providers to use content negotiation. Negotiation between the requesting agent and the server determines which representation is served (usually with the goal of serving the "best" representation a receiving agent can process). Figure 1.5 shows an example of the generated HTML page by the dereferencing mechanism associated with TransformON concept URIs.

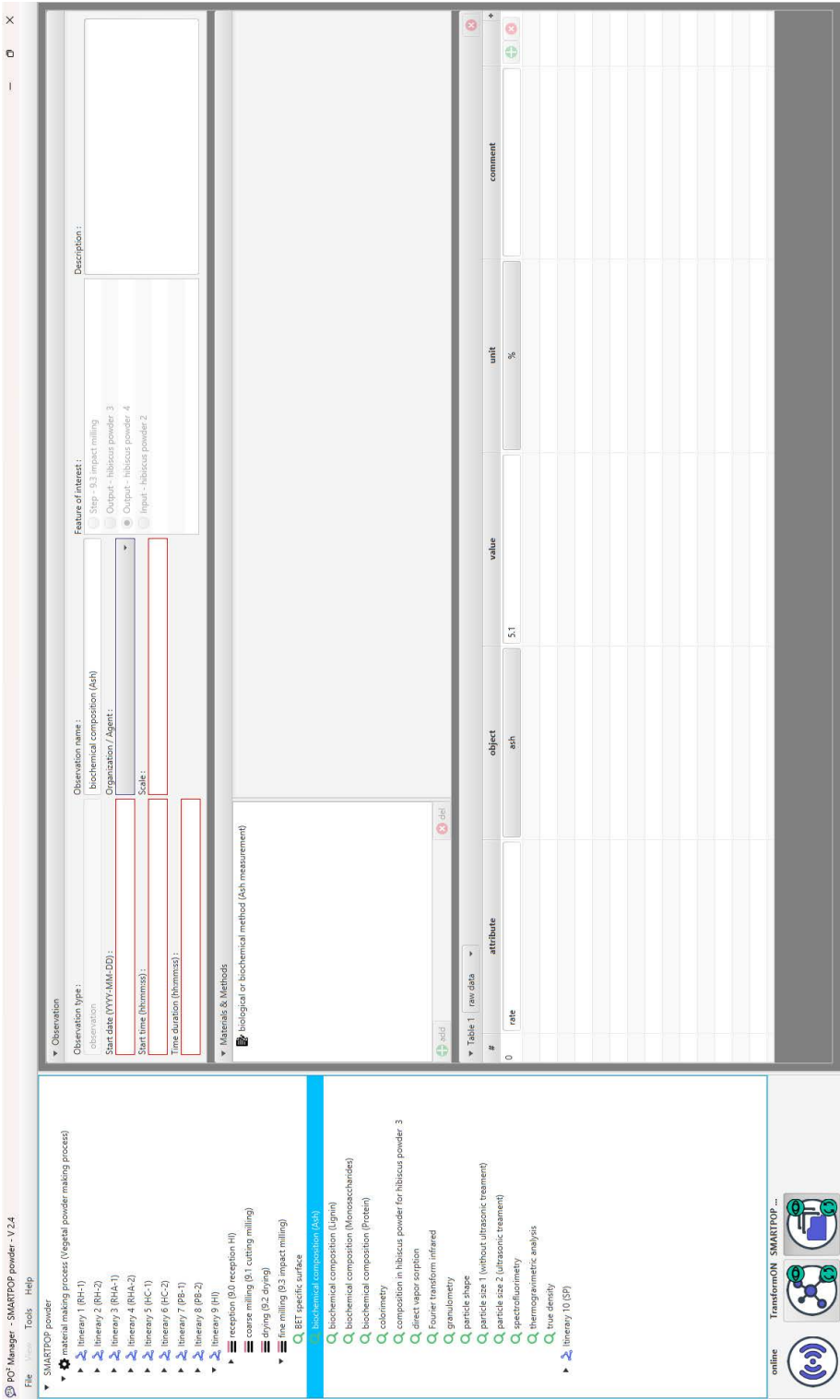


Figure 1.3. Screenshot showing process editing with PO² Manager: a detailed look at observations associated with hibiscus powder (ash content in Table 1).

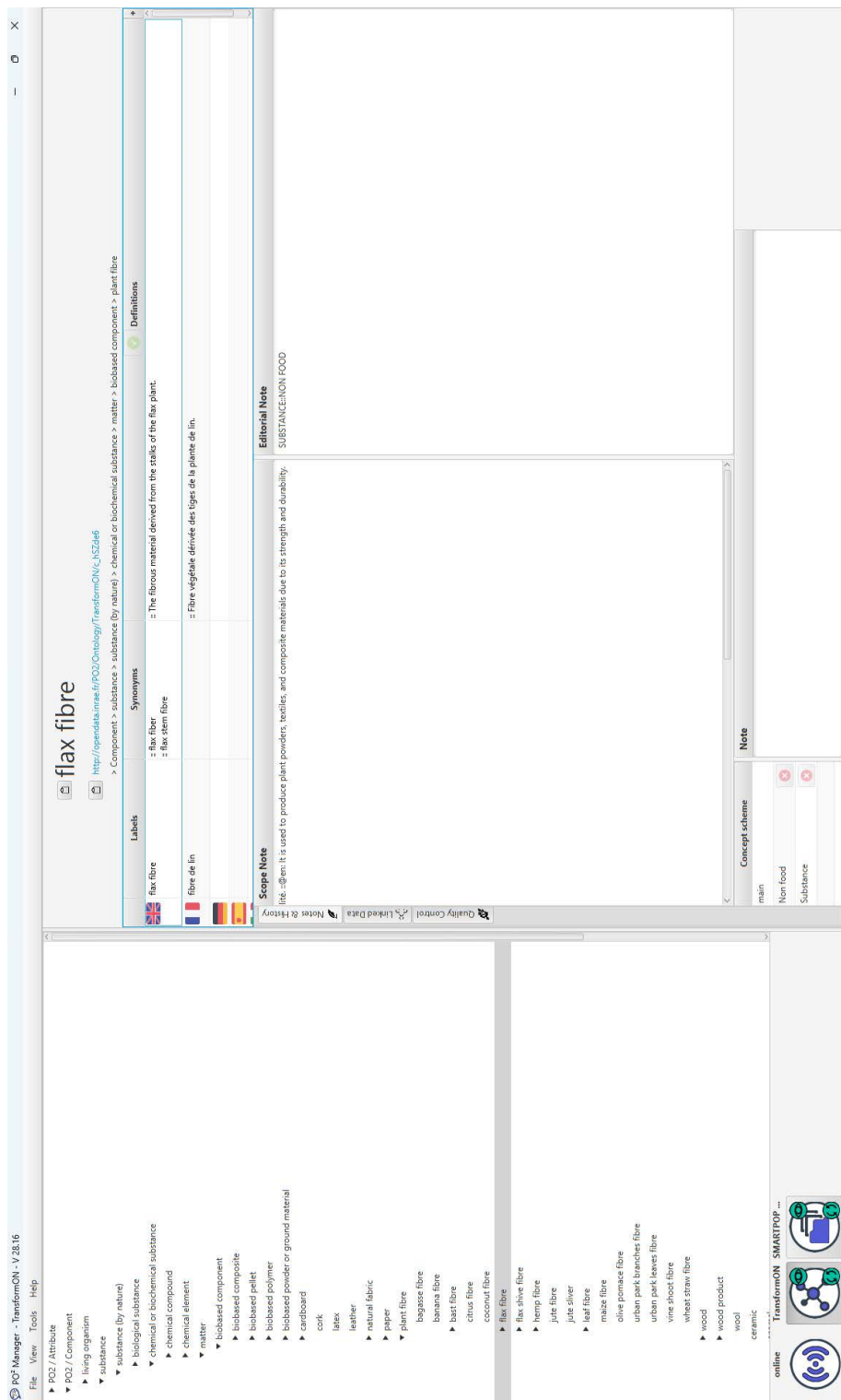


Figure 1.4. A screenshot from PO² Manager: an excerpt of TransformON on the left side of the window, and the definition of the concept “flax fiber” on the right side.

The screenshot shows the 'lodview' interface for the concept 'flax fibre'. The header includes the URL 'http://opendata.inrae.fr/PO2/Ontology/TransformON/c_hSZde6' and the type 'AN ENTITY OF TYPE: Class'. The main content area displays various properties of the concept:

- skos:definition**: EN FR. The fibrous material derived from the stalks of the flax plant. rdf:langString
- <http://opendata.inrae.fr/PO2/core/definitionStatus>**: validated. xsd:string
- skos:altLabel**: flax fiber @en (rdf:langString), flax stem fibre @en (rdf:langString)
- skos:editorialNote**: SUBSTANCE. xsd:string
- skos:note**: xsd:string
- skos:prefLabel**: DE EN ES FR IT. flax fibre (rdf:langString)
- skos:scopeNote**: @fr: Elle est utilisée pour produire des poudres végétales, des textiles, et des matériaux composites en raison de sa résistance et de sa durabilité. @en: It is used to produce plant powders, textiles, and composite materials due to its strength and durability. xsd:string

Figure 1.5. The HTML page that is generated by the dereferencing mechanism associated with TransformON concept URI for the concept “flax fiber”.

To find a concept, the user can navigate through the hierarchy or use the Search function in PO² Manager. The Search box is shown in Figure 1.6 below, with the different search results of the searched term displayed on the left side, as well as the location of the searched concept in a hierarchical view on the right side. The user can select one or several “branches” (thesauri) corresponding to each of the core concepts (Attribute, Component, Material, Method, Process, Scale, Step).

1.2.2 Solutions proposed to match vocabularies defined in different data sources for integration purposes

As noted in Section 1.1.1, additional external data must be found to enrich the datasets. Two workflows are presented to solve the second research question Q1.2, depending on the alignment context:

- *Unsupervised method (OntoMapF)*: No alignment examples between the data sources are available to map: in this case, a workflow based on a pivot ontology and background knowledge is proposed;

- *Supervised method (Oqlassif)*: Numerous alignment examples between data sources are available, and a pivot ontology (a workflow based on a machine learning approach) is proposed.

The OntoMapF method to compute the alignment presented in Buche et al. (2021a) aligns a food composition database (FCDB) with a pivot ontology (specifically, the SIREN part of FoodOn that contains LanguaL™ codes) as shown in Figure 1.7a.

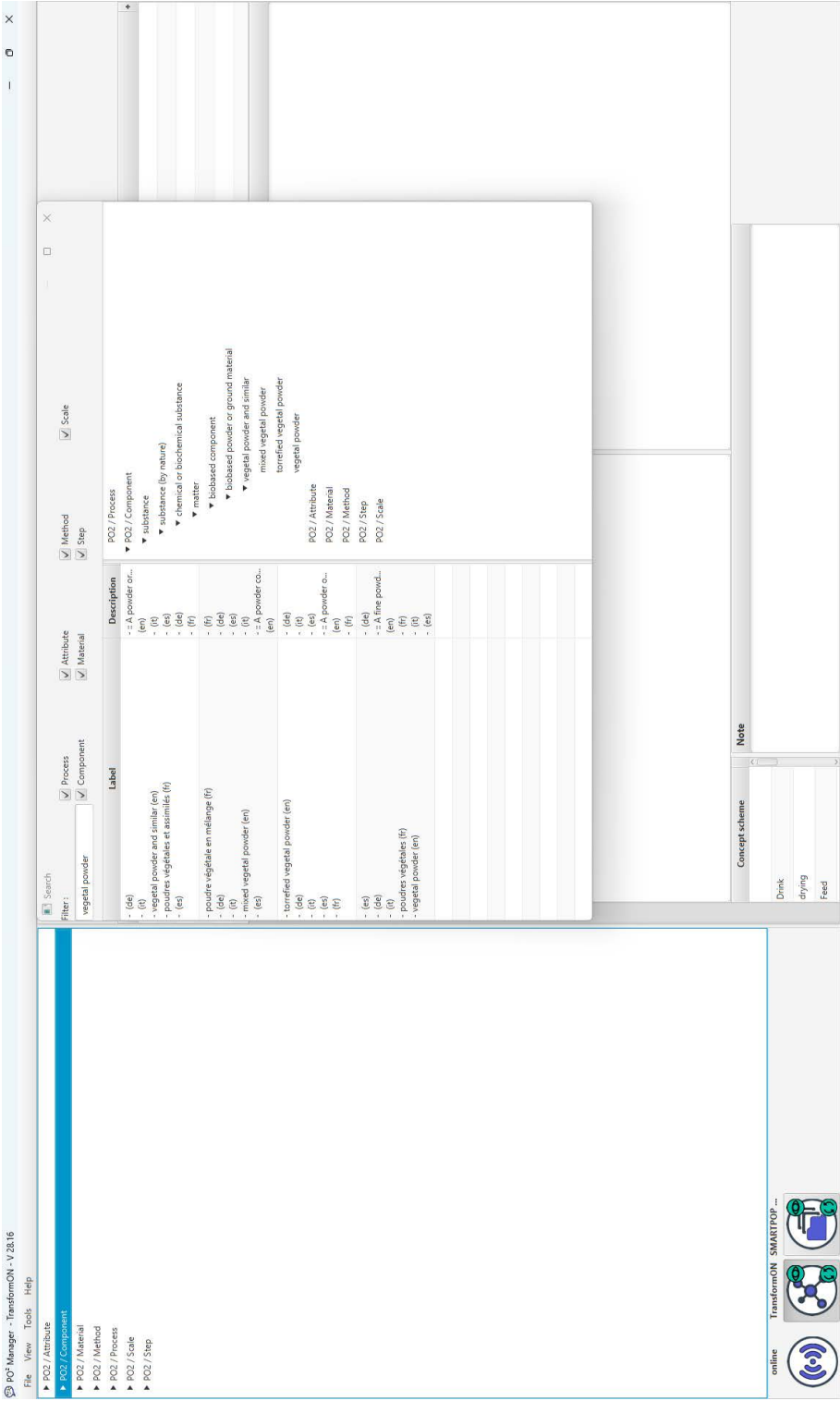


Figure 1.6. PO² Manager screenshot showing the navigation through the hierarchy and use of the Search function to find concepts in TransformON.

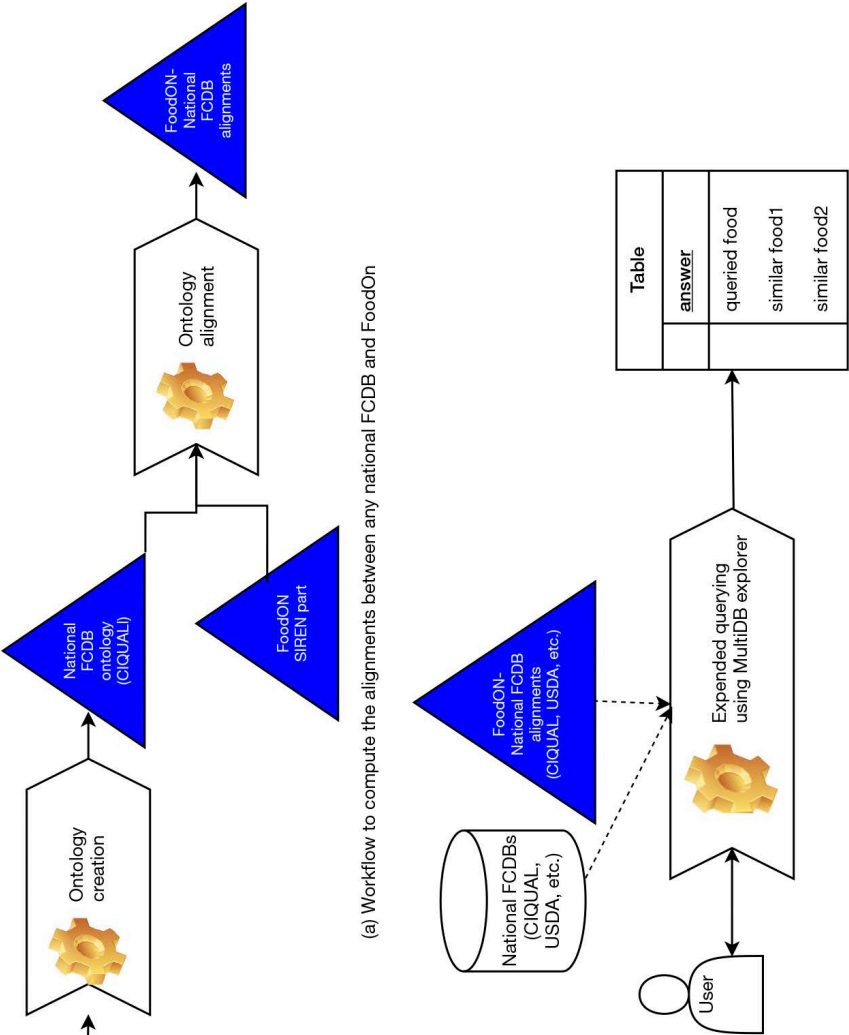


Figure 1.7. Workflows associated with the unsupervised alignment method.

This method is based on the EuroFIR interchange system (Pakkala et al., 2010), which harmonizes FCDB database structures. The method makes the assumption that the English food terminology associated with the FCDB has been described using the LanguaL™ thesaurus. LanguaL™ is based on the following hypotheses: (i) any food can be systematically described using a combination of characteristics; (ii) these characteristics can be categorized into viewpoints, composed of one or several facets, and coded for computer processing; and (iii) the resulting food description codes can be used to retrieve food data from FCDBs. Table 1.1 shows the 14 facets of LanguaL™.

Table 1.1 LanguaL™ characteristics and associated facets reproduced from Buche et al. (2021a)

CHARACTERISTIC	FACET
FOOD GROUP	A. Product type including Codex alimentarius classification for food and feeds and other international classifications
FOOD ORIGIN	B. Food source species of plant or animal, or chemical food source C. Part of plant or animal
PHYSICAL ATTRIBUTES	E. Physical state, shape, or form
PROCESSING	F. Extent of heat treatment G. Cooking method H. Treatment applied
PACKAGING	J. Preservation method K. Packing medium M. Container or wrapping N. Food contact
DIETARY USES	P. Consumer group/dietary use/label claim
GEOGRAPHICAL ORIGIN	R. Geographical places and regions
MISCELLANEOUS CHARACTERISTICS	Z. Adjunct characteristics of food

Thanks to LanguaL™, a food may be described using a combination of descriptors regrouped into facets. Each descriptor is defined by a standard controlled term and a code. Each standard controlled term belongs to a specialization hierarchy that is defined for each facet.

Table 1.2 shows that the food *Cooked pork shoulder, choice* from the CIQUAL database is described with 14 descriptors (i.e., a code associated with a controlled term, regrouped into facets labeled A, B, C, E, F, G, H, J, P, and Z).

Table 1.2 LanguaL™ descriptors for the food *Cooked pork shoulder, choice* used in the CIQUAL database reproduced from Buche et al. (2021a)

LANGUAL™ CODE	LANGUAL™ TERM
A0279	CURED MEAT (US CFR)
A0797	PRESERVED MEAT (EUROFIR)
B1136	SWINE

LANGUAL™ CODE	LANGUAL™ TERM
C0270	SKELETAL MEAT PART, WITHOUT BONE AND SKIN, WITHOUT SEPARABLE FAT
E0137	SLICED
F0014	FULLY HEAT-TREATED
G0003	COOKING METHOD NOT APPLICABLE
H0253	CURED OR AGED
H0367	SALT ADDED
J0100	PRESERVED BY ADDING CHEMICALS
J0131	PRESERVED BY CHILLING
P0024	HUMAN CONSUMER, NO AGE SPECIFICATION
Z0010	CHOICE GRAD
Z0043	SHOULDER (MEAT CUT)

The alignment method leverages both the food descriptions within the LanguaL™ thesaurus and the English terminology used in the FCDB. It consists of two primary steps:

- Transformation: Transforming the FCDB English terminology and the associated LanguaL™ description into an FCDB ontological format.
- Alignment computation: Calculating the alignment scores between food concepts based on information that is both semantic (food descriptions within the LanguaL™ thesaurus) and syntactic (English terminology).

The example below is an excerpt of the OWL definition, expressed in the Turtle syntax, associated with the concept *Cooked pork shoulder, choice* corresponding to the same term in CIQUAL shown in Table 1.2. The excerpt is restricted to LanguaL™ codes A0279 and B1136:

```
@prefix: <#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

### Base classes
:FoodConcept a owl:Class .

### Facet A – Product type
:A_PRODUCT_TYPE a owl:Class ;
  rdfs:subClassOf :FoodConcept .

:CURED_MEAT_US_CFR a owl:Class ;
  rdfs:subClassOf :A_PRODUCT_TYPE .

### Facet B – Food source
:B_FOOD_SOURCE a owl:Class ;
  rdfs:subClassOf :FoodConcept .

:SWINE a owl:Class ;
  rdfs:subClassOf :B_FOOD_SOURCE .
```

```

### Property
:has_Facet a owl:ObjectProperty ;
  rdfs:domain :FoodConcept ;
  rdfs:range :FoodConcept .

### Concept described by facets
:Cooked_pork_shoulder_choice a owl:Class ;
  rdfs:label "Cooked pork shoulder, choice"@en ;
  rdfs:subClassOf :FoodConcept ,
    [ a owl:Restriction ;
      owl:onProperty :has_Facet ;
      owl:someValuesFrom :CURED_MEAT_US_CFR
    ] ,
    [ a owl:Restriction ;
      owl:onProperty :has_Facet ;
      owl:someValuesFrom :SWINE
    ] .

```

This approach offers a significant advantage when aligning a new FCDB with FoodOn. By leveraging previously aligned FCDBs, it enables automatic alignment of the new FCDB through a transitive property, eliminating the need for manual bilateral alignments between individual FCDBs. For example, in Figure 1.8, the CIQUAL concept *Courgettes, puree* (resp. USDA *squash, Indian, raw (Navajo)*) has been aligned with the FoodOn concept *squash vegetable food product* using a *kind_of* relation. Since they both belong to the same FoodOn category, the respective composition tables may be considered as close.

Finally, a new application called MultiDB explorer was created during the French national Meatylab project involving both industrial and academic partners. MultiDB explorer currently integrates several national FCDBs in a unique data warehouse, including CIQUAL and the USDA SR Legacy Database. As shown in Figure 1.7b, MultiDB explorer can be used to simultaneously query all the FCDBs stored in the data warehouse using the terminology of a given FCDB (e.g., CIQUAL) in the national language or in English. The result is extended to similar concepts (see Figure 1.8) found in other FCDBs (e.g., those from the USDA). See Buche et al. (2021a) for more details.

Oqlassif supervised alignment method workflow

When a relatively large number of alignments between heterogeneous data sources are available, we propose using a supervised machine learning approach.

The Oqlassif method was tested with a training dataset containing 60,400 food products, classified into 32 sectors and more than 643 families. This dataset was produced by the French food observatory team (OQALI¹⁸). The OQALI coding system and associated database include much of the information available on the packaging of processed food products on the French market (nutritional composition, ingredients, etc.). It is structured into food categories, with each collected on a regular basis to follow changes in the offer over time. This means that not all food categories are collected the same year, and the information made available corresponds to the products available on the market at a time *t*.

18. Observatoire de l'alimentation (<https://www.oqali.fr/>)

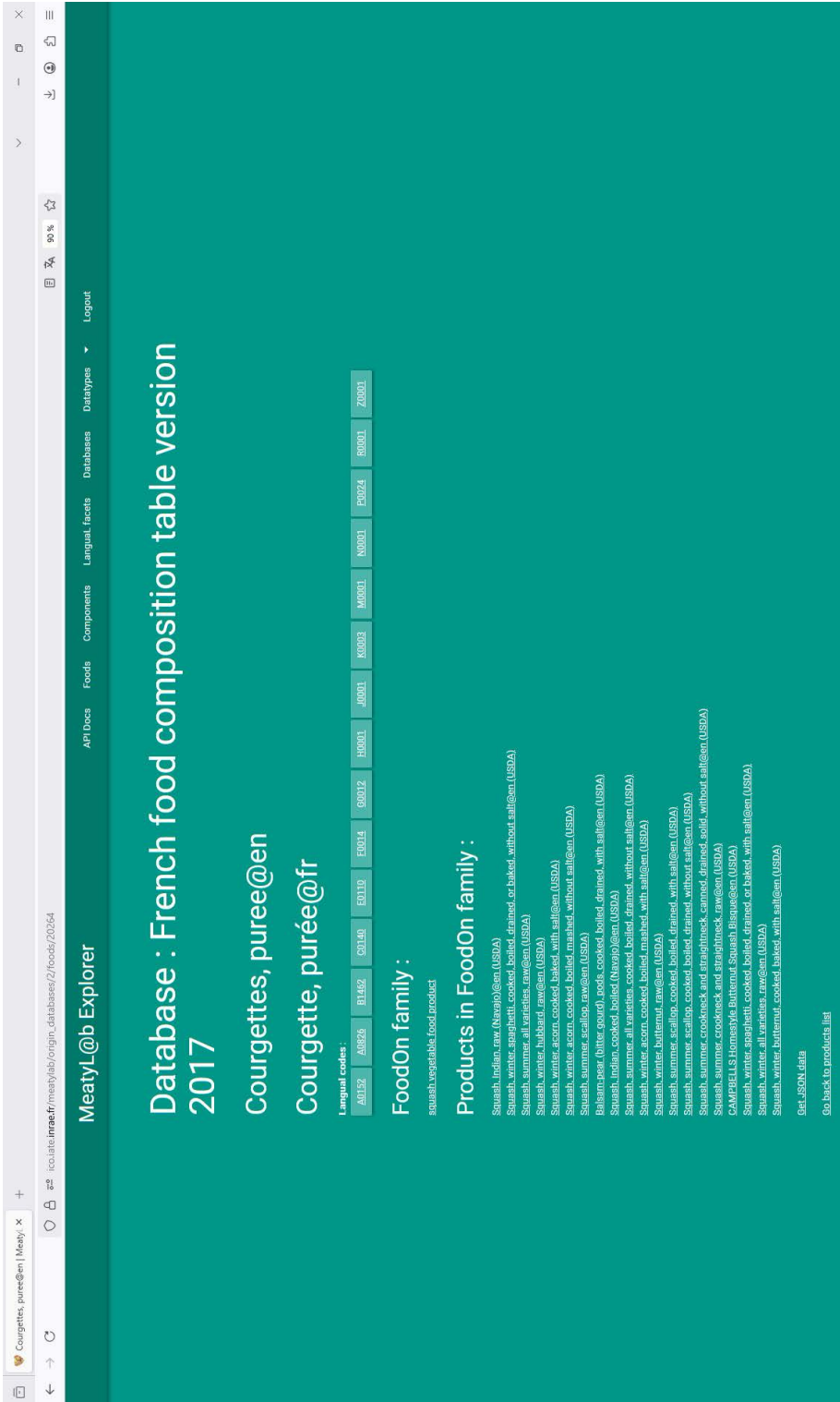


Figure 1.8. Similar USDA concepts for the CIQUAL food Courgettes, puree generated by the MultiDB explorer tool presented in (Buche et al., 2021a).

The 60,400 food products belonging to the OQALI food categories are described in French with four textual variables: sale description name, description, product detail (flavor, etc.), and storage method (see Table 1.3). The four textual variables are selected by the OQALI experts to maintain categorization when the OQALI coding system is updated. A given product is also categorized according to a sector and family. For example, “Dairy products and fresh desserts” is one of the largest sectors and contains several families, such as “Classic yogurts and sweetened fermented milks” and “Classic sweet fresh cheeses,” among others.

For the benefit of English-speaking readers, each table is first presented in French, followed by its English translation. In the current implementation of the algorithm, preprocessing is performed on French terms.

Table 1.3a Examples of textual variables associated with a product description

Dénomination de vente	Secteur	Famille	Nom du produit	Précision	Mode de conservation
XXX*	Produits laitiers et desserts frais	Yaourts et laits fermentés sucrés classiques YaourtY	Lait fermenté sucré à boire aux fruits enrichi en vitamines b2 b12 d	Framboise	Frais
YYY* PARFUM FRAISE	Produits laitiers et desserts frais	Yaourts et laits fermentés sucrés classiques YaourtY	yaourt à boire sucré aromatisé	Fraise	Frais

*Brand names were anonymized.

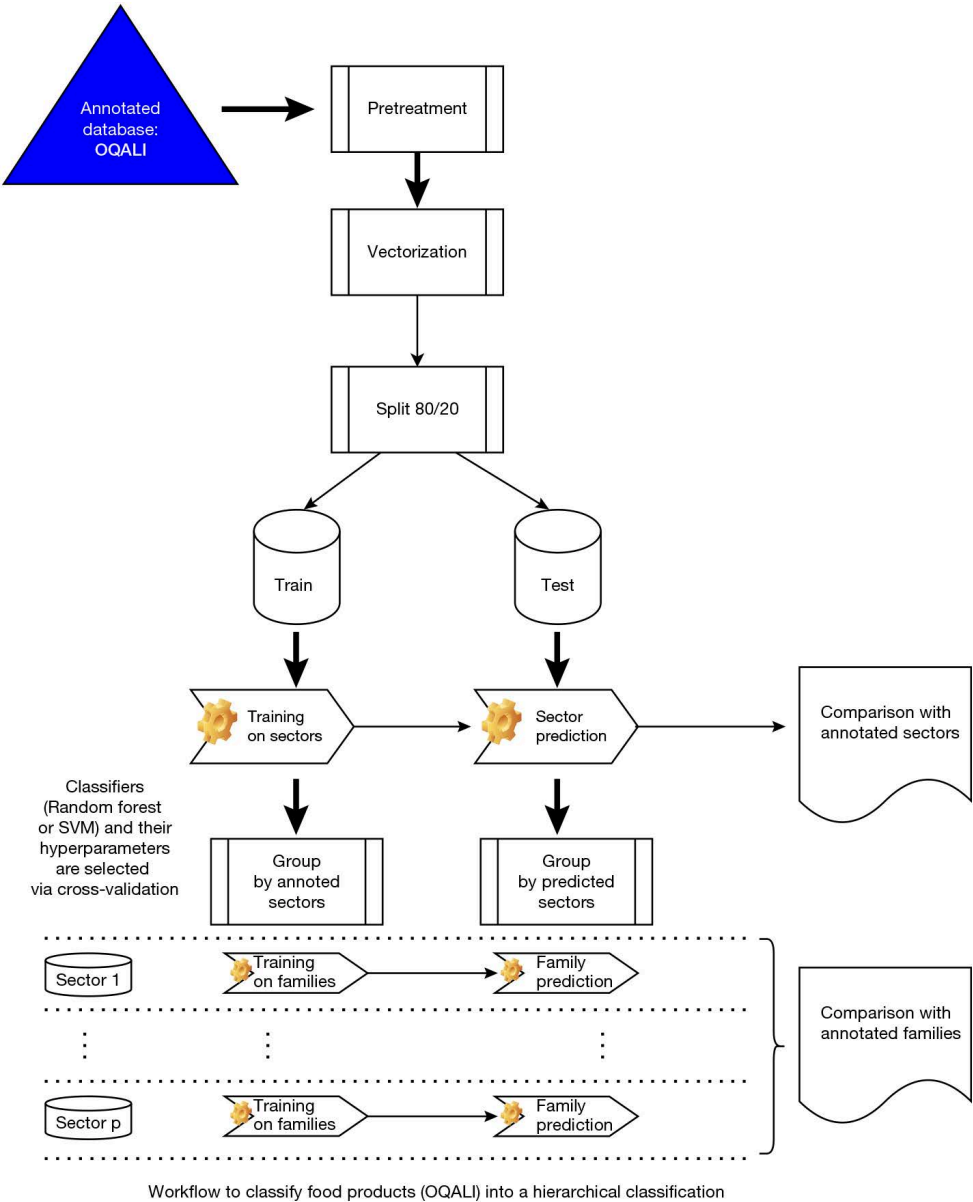
Table 1.3b Translation of Table 1.3a

Sale_description	Sector	Family	Description	Product_detail	Storage_method
XXX*	Dairy products and fresh desserts	Classic yogurts and sweetened fermented milk	sweetened fermented milk drink with fruit and vitamins b2, b12, and d	Raspberry	Refrigerated
YYY* STRAWBERRY FLAVOR	Dairy products and fresh desserts	Classic yogurts and sweetened fermented milk	sweetened flavored drinking yogurt	Strawberry	Refrigerated

*Brand names were anonymized.

Some 60,000 alignment examples between data sources and the OQALI coding system are available. To automatically classify/align a new example, a workflow based on a machine learning approach is proposed and presented in Figure 1.9.

Figure 1.9 shows that the food product descriptions are pretreated (see Tables 1.4 to 1.7). The pretreated food descriptions are then transformed into vectors (see Table 1.8). The training step is divided into two parts. Data are split into one training set (80%) and one test set (20%).



Workflow to classify food products (OQALI) into a hierarchical classification

Figure 1.9. The workflow designed to create the Oqlassif’s classifiers using the “cascade” method.

Table 1.4a Special characters, extra spaces, uppercase letters, and accents are removed, but numbers are kept

Dénomination de vente	Nom du produit	Précision	Mode de conservation
XXX*	lait fermente sucre a boire aux fruits enrichi en vitamines b2 b12 d	framboise	Frais
YYY* parfum fraise	yaourt a boire sucre aromatise	fraise	Frais

*Brand names were anonymized.

Table 1.4b Translation of Table 1.4a

Sale_description	Description	Product_detail	Storage_method
XXX*	sweetened fermented milk drink with fruit and b2 b12 d vitamins	raspberry	Refrigerated
YYY* strawberry flavor	sweetened flavored drinking yogurt	strawberry	Refrigerated

*Brand names were anonymized.

Table 1.5a French stop words (le, la, les, à, etc.) are removed

Dénomination de vente	Nom du produit	Précision	Mode de conservation
XXX*	lait fermente sucre boire fruits enrichi vitamines	framboise	Frais
YYY* parfum fraise	yaourt boire sucre aromatise	fraise	Frais

*Brand names were anonymized.

Table 1.5b Translation of Table 1.5a

Sale_description	Description	Product_detail	Storage_method
XXX*	sweetened fermented milk drink fruit b2 b12 vitamins	raspberry	Refrigerated
YYY* strawberry flavor	sweetened flavored drinking yogurt	strawberry	Refrigerated

*Brand names were anonymized.

Table 1.6a Lemmatization: plurals are made singular, conjugated forms are changed to infinitives, etc.

Dénomination de vente	Nom du produit	Précision	Mode de conservation
XXX*	lait fermenter sucrer boire fruit enrichir	framboise	Frais
YYY* parfum fraise	yaourt boire sucre aromatiser	fraise	Frais

*Brand names were anonymized.

Table 1.6b Translation of Table 1.6a

Sale_description	Description	Product_detail	Storage_method
XXX*	sweeten ferment milk drink fruit b2 b12 vitamins	raspberry	Refrigerated
YYY* strawberry flavor	sweeten flavor drinking yogurt	strawberry	Refrigerated

*Brand names were anonymized.

Table 1.7a Stemming: the word roots are identified

Dénomination de vente	Nom du produit	Précision	Mode de conservation
XXX*	lait ferment sucr boir fruit enrichi	framboise	Frais
YYY* parfum fraise	yaourt boir sucr aromatis	fraise	Frais

*Brand names were anonymized.

Table 1.7b Translation of Table 1.7a

Sale_description	Description	Product_detail	Storage_method
XXX*	sweeten ferment milk drink fruit b2 b12 vitamin	raspberry	Refrigerated
YYY* strawberry flavor	sweeten flavor drink yogurt	strawberry	Refrigerated

*Brand names were anonymized.

If a column containing the sale description name is provided, the brand, when it appears in the variables, is considered as a name entity and is not pretreated. It is not necessary to pretreat the storage_method variable, because there are only three possible cases: refrigerated, frozen, ambient. Those product descriptions are then put together in the same variable and transformed into binary vectors: for each word (feature) present in the corpus, 1 means the product description contains it, 0 that it does not. The dimension of the vectors is determined by the number of pretreated words available in the training dataset.

Table 1.8a Vectorization of the food product descriptions (the vector is split into two parts)

XXX*	lait	ferment	sucr	boir	fruit	enrichi	vitamin	b2	b12
1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	0	0	0	0	0
framb	YYY*	parfum	frais	yaourt	aromatis	Frais	Ambiant	Surgelé	
1	0	0	0	0	1	1	0	0	
0	1	1	1	1	1	1	0	0	

Table 1.8b Translation of Table 1.8a

XXX*	sweeten	ferment	milk	drink	fruit	b2	b12	d	vitamin
1	1	1	1	1	1	1	1	1	1
0	0	0	1	1	0	0	0	0	0
raspberry	YYY*	strawberry	flavor	yogurt	Refrigerated	Ambient	Frozen		
1	0	0	0	0	1	0	0		
0	1	1	1	1	1	0	0		

*Brand names were anonymized.

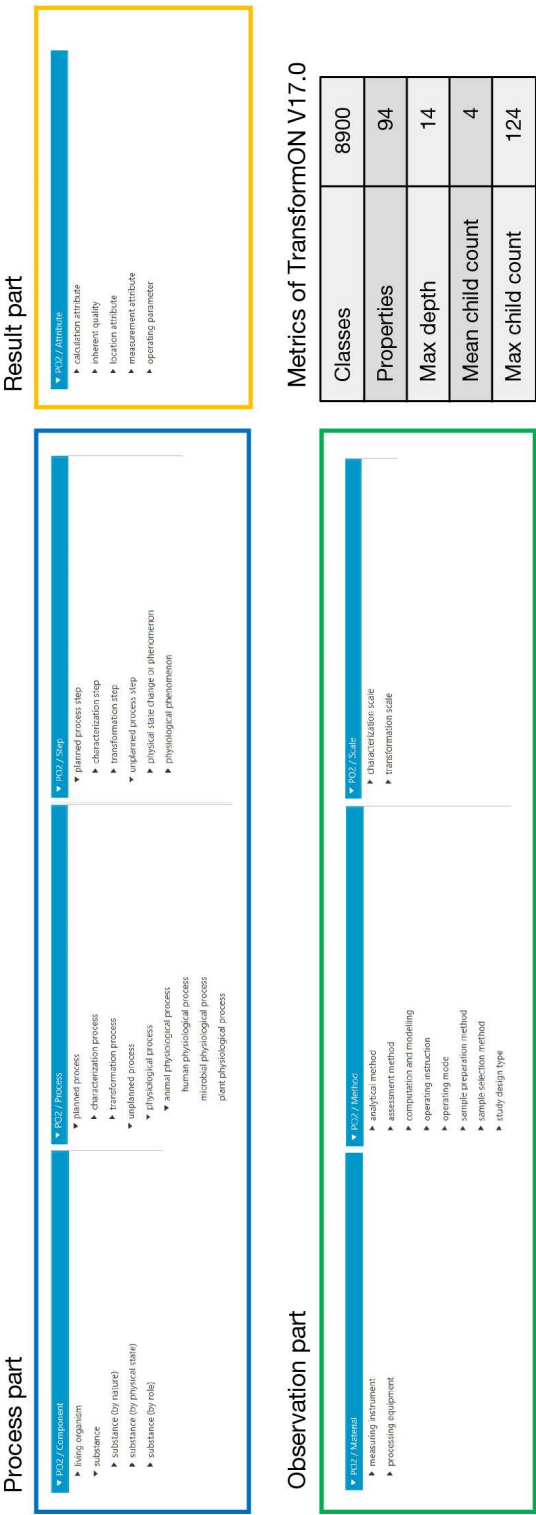
Cross-validation is then used (see Figure 1.9) to select hyperparameters for chosen classifiers (random forest and SVM). More precisely, for each set of hyperparameters, five different training/test subpartitions are made from the whole training set. A classifier with those parameters is trained on the training set of each subpartition, then tested on the test set of the subpartition. The mean of the five different accuracy values is used to evaluate the performance. One classifier is selected using the cross-validated performance indicators (precision, recall, and F1-score) presented in section 1.3.2 below to be trained on the whole training set in order to classify products into sectors. This classifier is used to predict sectors in the whole test set. The same procedure is then used for each of the 30 sectors to fine-tune the hyperparameters associated with family classifiers belonging to the considered sector. This “cascade” method helps reduce execution times (the more features a dataset has, the more times it takes to train/predict). For example, using our server, execution times to train all the classifiers decreased from dozens of hours to only four. It still takes four days to test and select hyperparameters for the sector classifier, but preselected hyperparameters can be used for training to save time.

►► 1.3 Outcomes

1.3.1 TransformON metrics

Figure 1.10 shows the upper levels of the specialization hierarchies of TransformON domain ontology based on the PO² core model. Under Process part, users will find the PO²/Component, PO²/Process, and PO²/Step specialization hierarchies; under Observation part the PO²/Material, PO²/Method, and PO²/Scale; and under Result part the PO²/Attribute specialization hierarchy and the unified code for units of measure (UCUM), a coding system designed to include all current units of measure commonly used in international science, engineering, and business.

In constructing the PO²/Component branch, we distinguished between living organisms and substances such as various types of biomass, energy, chemical compounds, etc. Living organisms are the biological sources of raw materials used in processing after harvesting or slaughtering. The “living organisms” class groups together categories of living organisms according to the three-domain taxonomic classification system (Archaea, Bacteria, Eukarya). Substances can be described more precisely based on three characteristics: their nature (biological or physicochemical), their physical state (solid, liquid, etc.), and their role (food or nonfood use).



The “Dietary substance” hierarchy is divided into three subclasses: “Food,” “Feed,” and “Dietary constituents” (Figure 1.11). The “Food” subclass is based on the main hierarchy proposed in FoodEx2. It encompasses a total of 4,242 classes, directly imported from FoodEx2. To facilitate future alignments with FoodOn, certain intermediate classes have been added. These are used, for example, to distinguish vertebrate animals from invertebrates, or animal substances such as raw milk or eggs. The “Feed” branch includes 759 classes imported from FoodEx2, but its hierarchical structure has been modified. In fact, two parent subclasses have been created in the hierarchy: one to group all primary sources of animal feed, and the other to group compound feed. The latter follows the same logic as the “Food” branch. Food constituents corresponding to nutritional elements are listed in the “Dietary constituents” branch, which comprises 49 classes. Dietary constituents will be aligned with the Compositional Dietary Nutrition Ontology (CDNO) ontology and with CheBi.



Figure 1.11. Details of the specialization hierarchies where the first three upper levels are shown.

The “Nonfood substance” hierarchy was created in TransformON to group together all substances used or produced during processing or characterization processes, taking into account their role or use: products and reagents involved in chemical reactions, enzymes, packaging materials, refrigerants, pollutant emissions, energy resources, and recyclable waste. It should be noted that the construction of the “recyclable waste” branch was based on existing references, such as the European Waste Catalogue (document annexed to Commission Decision 2000/532/EC).

Two types of processes have been defined in the TransformON ontology. Planned processes are controllable, in the sense that they follow procedures or experimental designs. In contrast, unplanned processes, such as biological or physiological processes (e.g., fermentation or human digestion), are observable phenomena. While not fully controllable, these processes can be influenced by manipulating control parameters. PO²/Steps are defined as the elements that make up a process. These can be changes of state or physical phenomena, biological or physiological process steps, characterization steps, or processing steps. These four classes or categories have been subdivided into finer subclasses that specialize the steps according to the type of action or phenomenon they represent.

Finally, under Observation, users will find three hierarchies: PO²/Material, PO²/Method, and PO²/Scale. The PO²/Material concepts can be categorized as either transformation or characterization equipment. The concepts in the PO²/Method hierarchy represent procedures, operating modes, or analytical protocols. PO²/Scale concepts can be used to indicate the scale of the object observed (measurement scale) or the scale at which the process is carried out (processing scale). The PO²/Attribute hierarchy groups together observable characteristics that can be measured or calculated. The hierarchy is divided into four parts: (i) calculation results (characteristics whose values are calculated from other values, such as index values or scores), (ii) intrinsic properties (characteristics inherent in objects), (iii) measurement attributes (characteristics obtained directly from a measurement), and (iv) operating parameters (characteristics used to control processing or characterization processes).

1.3.2 OntoMapF unsupervised alignment method assessment

The MultiDB explorer was used to address the absence of values in the CIQUAL FCDB for three specific nutrients (vitamin C, vitamin B12, and iron) identified by the industrial partners in the Meatylab project. Using the alignments provided by the OntoMapF process, the missing values were replaced with those linked to comparable foods in the USDA FCDB. This task was performed on 99 foods for which at least one of the values associated with the three nutrients was not known in CIQUAL FCDB and where at least one similar food could be automatically found in USDA FCDB. A total of 76 alignments were considered relevant by domain experts. For those 76 relevant alignments, 91% of unknown values in CIQUAL were enriched with values from USDA and 96% of known values in CIQUAL were completed with values from USDA. More details on the assessment may be found in Buche et al. (2021a).

Further experiments were performed to assess the unsupervised alignment method. A set of 181 food product concepts from CIQUAL, Gold Standard (GS) were manually linked to a concept in the FoodOn ontology with the “is-kind of” relation. Assessments presented in Table 1.9 concern “is-kind of” matches found with the combination of the

syntactic and the semantic scores, the syntactic score alone, and the semantic score alone for GS. The metrics used for assessment (precision, recall, and F1-score) were widely applied to evaluate information retrieval systems (Salton and McGill, 1986):

- Precision: percentage of products predicted in this class that truly belong to it;
- Recall: percentage of products belonging to this class that were predicted in it;
- F1-score: harmonic mean of precision and recall.

For a given food concept C in CIQUAL, if the best match belongs to the family associated with C in GS, it is considered as an “is-kind of” first match (Table 1.9). The best result in terms of F-score is obtained for the “is-kind of” first match with the semantic score alone.

Table 1.9 Family matches results (precision (P), recall (R), F1-score (F-M)) depending on a threshold with the CIQUAL Gold Standard [adapted from Buche et al. (2021a)]

Alignment method	First			
	Threshold	P	R	F-M
Combination Syntactic score	0.6	0.63	0.55	0.59
+ Semantic score	0.9	0.81	0.12	0.21
Syntactic score	0.6	0.75	0.24	0.36
Semantic score	0.6	0.81	0.63	0.71

1.3.3 Oqlassif supervised alignment method assessment

A set of 60,400 products were split into a training set (80%) and a test set (20%). The accuracy (a metric that measures how often the model correctly predicts the outcome) of sector classification is 99.2%, while families have an accuracy of 93.4%. When predicting multiple families, the accuracy score rises to 97.5% with only three families.

To evaluate the model’s performance, classification reports for the sector prediction and family predictions are made. For each class (sector or family), precision, recall, and F1-score are computed.

The performances associated with two methods are shown in Figure 1.12: the classic pipeline (Figure 1.12, top), which uses only product classifiers grouped into families, and the cascade pipeline (Figure 1.12, bottom), where product classifiers are first grouped into sectors and then into families belonging to the predicted sector. In the example shown in Figure 1.12, there are more families on the right side of the cascade pipeline’s graphic: this means that using a cascade pipeline helps some families improve their F1-score. Some families were still hard to predict because of their relatively small numbers. Users were then given manual tools to detect them.

During the training/test, the indicators below are generated:

- Confusion matrix, a table that compares the predicted labels of a classification model with the true labels. This matrix shows how many predictions were correct or incorrect for each class. Here, it shows where products are most likely to be mispredicted for each sector. In Figure 1.13, all products in the “Infant milks” sector are correctly predicted, so products can be found along the diagonal. This is not the case for products in the “Cereal bars” sector: between the 38 products truly belonging to the “Cereal bars with chocolate/chips/coating/filling” family, 30 are correctly

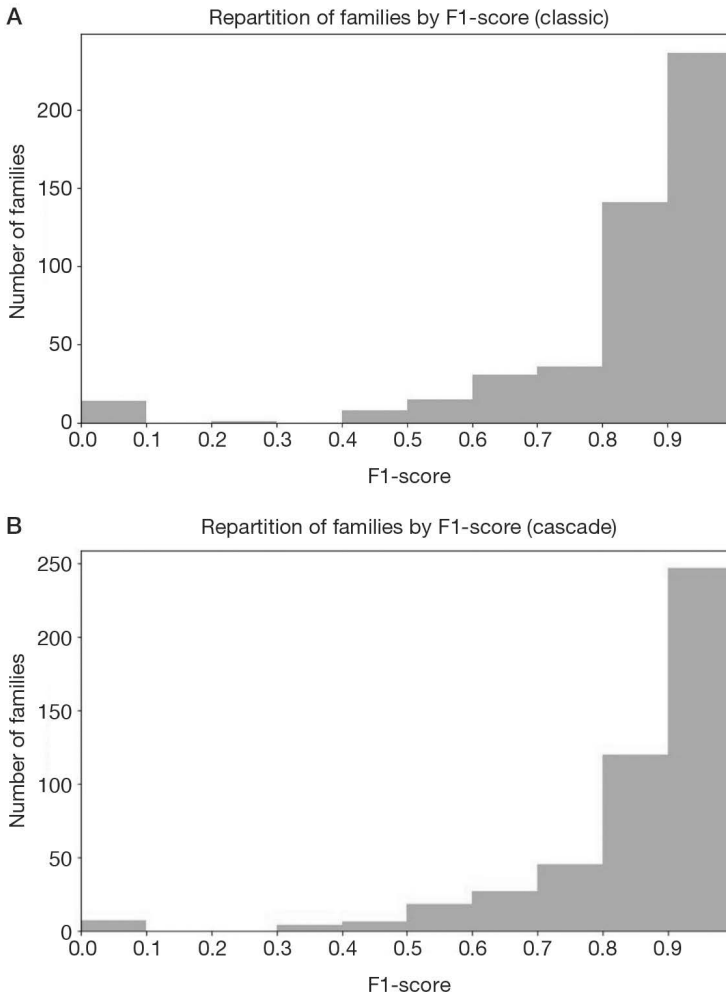


Figure 1.12. Number of families depending on their F1-score: classic pipeline (top) and cascade pipeline (bottom).

predicted, 6 are predicted in the “Cereal bars with nuts and chocolate” family, and 2 in the “Other cereal bars” family.

- Classification reports with precision, recall, and F1-score for each class.

When it comes to actual predictions:

- Multiple predictions: Three families are predicted for each product. This number was selected based on the OQALI expert team’s feedback. If the probability of the product belonging to the predicted sector or being the first predicted family are below 90% and 82%, respectively, a flag warns the users they should check the result. Figure 1.14 shows that the accuracy increases with the number of families to predict.

A desktop application was developed and is maintained by the OQALI team, who can train and select new hyperparameters for the classifiers. The team distributes the application to industrial partners.

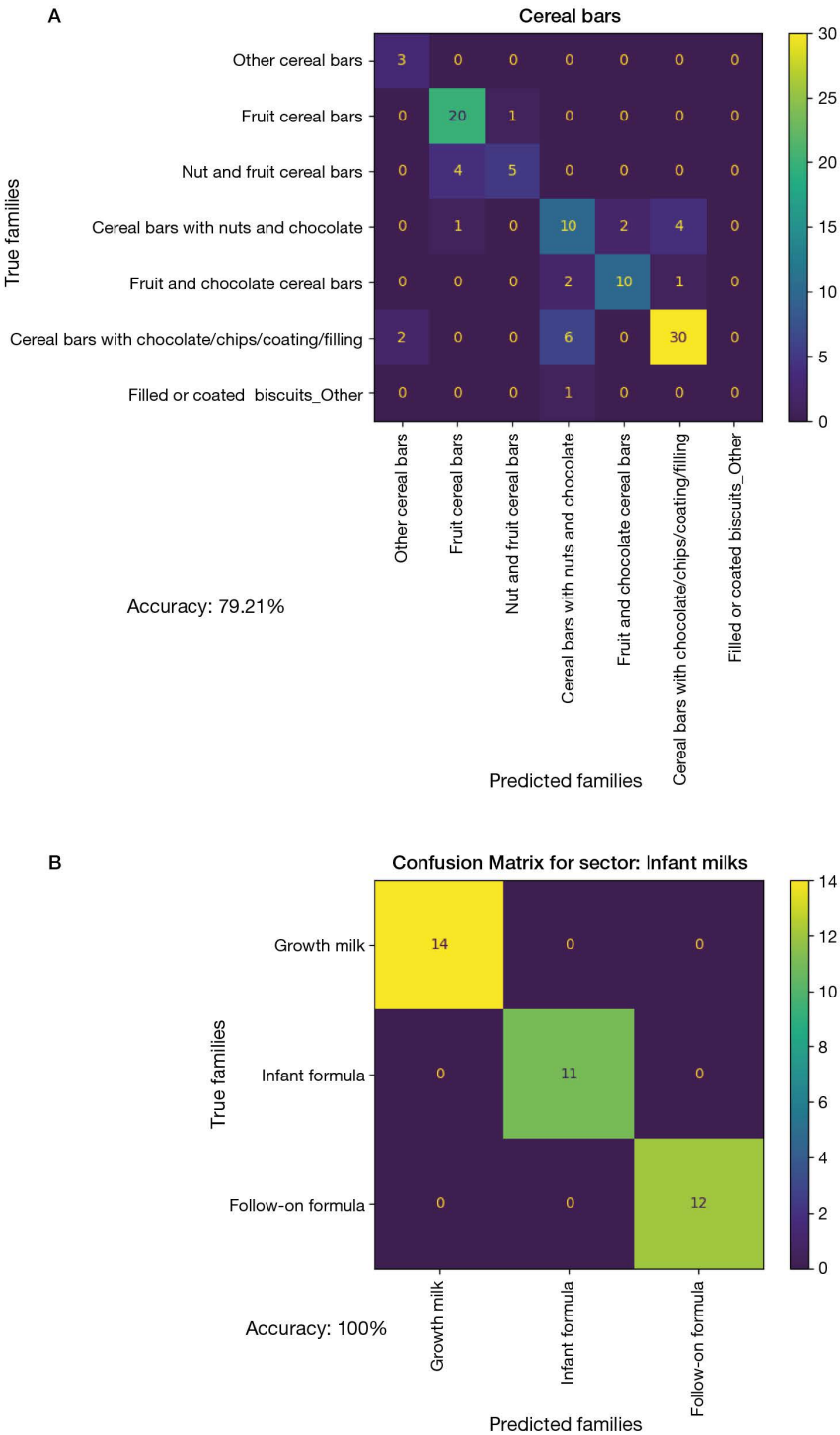


Figure 1.13. Two examples of a confusion matrix: “Cereal bars” (top) and “Infant milks” (bottom) sectors.

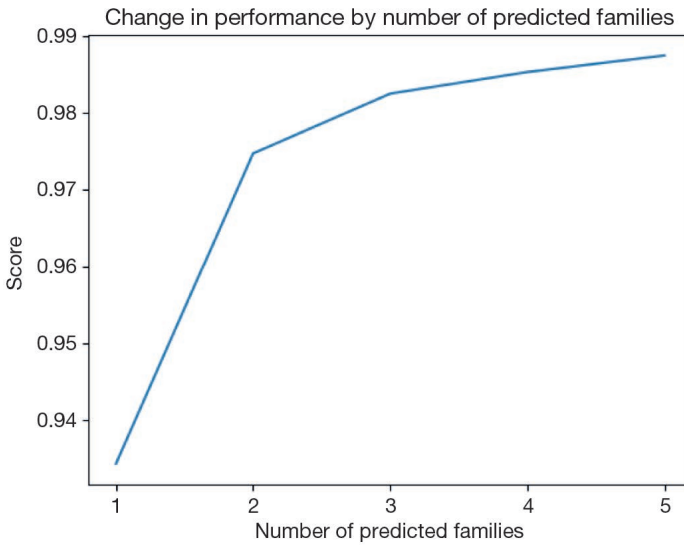


Figure 1.14. Change in accuracy according to the number of families to predict.

►► 1.4 Conclusion and resources

The two research questions covered in this chapter were:

- Q1.1: How can the multiple dimensions of the agrifood system be integrated to design sustainable, healthy, and affordable food and biobased products?
- Q1.2: How can vocabularies defined in different data sources be matched for integration purposes?

One solution, the TransformON domain ontology based on the PO² core model, was designed to answer Q1.1 (Section 1.2.1). Two solutions – unsupervised and supervised methods – were put forward in Section 1.2.2 to solve Q1.2 depending on the quantity of available alignments. These solutions have all been compared to the current state of the art. For each of them, the software workflow implementation was presented and illustrated. Available open resources (data, software, video, courses) are listed in Table 1.10. The main outcomes obtained using these solutions were also presented in Section 1.3. The TransformON domain ontology for Q1.1 and the supervised method for Q1.2 are used in production (Technology Readiness Level 9).

Ongoing research is now focusing on enriching TransformON with concept textual definitions using large language models (LLMs). Ultimately, the goal is to align TransformON with FoodOn for the food domain and with the INRAE Thesaurus for the agriculture, environment, and health domains (Aubin et al., 2024). This work will be carried out using the Simple Standard for Sharing Ontological Mappings (SSSOM) (Mapping Commons¹⁹; Matentzoglou et al., 2022), which is a community-driven standard designed to facilitate the exchange and integration of semantic entity mappings. As data interoperability becomes increasingly crucial across various

19. A Simple Standard for Sharing Ontology Mappings (SSSOM). GitHub. Retrieved from <https://mapping-commons.github.io/sssom/>

domains, SSSOM provides a standardized format to share mappings, enabling researchers and developers to more easily connect and utilize diverse datasets.

In Chapter 2, we will present the PO² task ontology, which has been specialized to obtain the domain ontology TransformON.

Table 1.10. Available resources

Research question	Proposed solution	Main scientific papers	Available resources			
			Data papers and datasets	Software codes	Videos	Courses and technical reports
Q1.1	PO2/ TransformON	Ontological model (Weber et al., 2023)	Ontological resource in OWL TRANSFORMON Summary AgroPortal see Section 2.3.1 Indicators for datasets	Ontology management in PO ² Manager: https://hal.science/hal-04313202	TransformON management with PO ² Manager: https://hal.inrae.fr/hal-05682889v1	
Q1.2a	Supervised alignment method	Ontological models and alignment method (Buche, et al., 2021a)	CIQUAL/Langual OWL ontology https://doi.org/10.15454/6CEYU3 SIREN/Langual OWL ontology https://doi.org/10.15454/5LLGVY CIQUAL/FoodOn Gold standard alignment https://doi.org/10.15454/BVXD71 CIQUAL/USDA Gold standard alignment https://doi.org/10.15454/4XIBS9	https://hal.inrae.fr/hal-05505080v1		
Q1.2b	Unsupervised alignment method		Food product descriptions including OQALI categorization: https://doi.org/10.57745/B4QHU8	Package: https://hal.science/hal-05507341 Interface: https://hal.science/hal-05507331		Technical report (in French): https://hal.inrae.fr/hal-05507173

Chapter 2

Task Ontologies for Product Processing and Characterization Processes

Liliana Ibanescu, Julien Couteaux, Julien Cufi, Stéphane Dervaux, Hervé Guillemain, Alrick Oudot, Emmanuele Piaud, Magalie Weber, Patrice Buche

► 2.1 Introduction

2.1.1 Issues and challenges

Food and biobased products manufacturers face significant challenges due to an aging workforce, the shift to Industry 4.0, and recruitment difficulties, particularly in sectors like the French agrifood industry, where many positions remain unfilled annually. Building intelligent digital platforms for acquiring, storing, and transferring knowledge is critical, especially as automation and data-driven decision-making rise. Full automation is unlikely, so humans will continue working in crucial roles, meaning that expertise and know-how must be preserved. Companies that manage product processing rely on their know-how to tailor their practices to the prevailing raw material variations, consumer expectations, and regulations, leaning heavily on on-the-job training. Industries must use digital tools to structure and exploit knowledge graphs and establish effective methodologies for knowledge transfer to ensure sustainability.

In the food and biobased products industry, a production process is composed of the steps through which raw materials (inputs) are transformed into one or more final products (outputs). Analyzing these processes is particularly challenging due to their heterogeneity (involving different variables and scales) and the time-dependent flow of information. To analyze or compare various processes, the measurements (or observations) that provide qualitative or quantitative values for a property associated with either an input/output or a process step/sequence of steps must be represented in a knowledge graph. Observable properties are also referred to as characteristics or attributes.

Production line management in factories is highly dependent on the operator's know-how and experience. Sharing collective expertise, which includes informal knowledge, is the best way to enhance overall quality. Informal knowledge is knowledge that has not been acquired during conventional or formal training, but rather through individual intentional or fortuitous experiences. The development of knowledge engineering methods to exploit knowledge graphs opens up new perspectives in terms of the preservation and data management of operational experience. These methods propose complex automatic reasoning techniques that go well beyond describing standard processes.

Today's Internet and the Web are much more than just semi-structured documents linked by hyperlinks. They also host a vast number of technical and scientific docu-

ments, on both the surface Web and the hidden Web (such as digital libraries), which often contain data tables. These tables can be considered as small relational databases, despite lacking the explicit metadata typical of databases. They represent a valuable external source that can be used to enrich internal knowledge graphs in companies or institutes focused on specific applications. Indeed, the integration of this data without a loss in consistency and meaning thanks to adapted extraction, interrogation, and machine learning mechanisms will allow the development of decision models based on reliable, credible, and traceable data. One of the first steps to integrate this external source of knowledge involves harmonizing external data with internal data, ensuring that the external data is expressed using the same vocabulary as that used for indexing the internal data. Once that is done, the reliability of external knowledge sources may be assessed before taking that information into account for decision support purposes.

The following research questions are addressed in this chapter:

- Q2.1: How can a process to turn raw materials into intermediate and final products be represented and include associated observations?
- Q2.2: How can the know-how that enables explanations and recommendations on technological actions to maintain/reach product quality or correct product quality defects be represented?
- Q2.3: How can data tables associated with external sources using the same vocabulary as that used for indexing the internal data be extracted and represented?
- Q2.4: How can the reliability of external data sources be assessed, and how can this assessment be accounted for in decision support systems?

2.1.2 State of the art and positioning, including the presentation of available scientific papers describing our proposal

As already mentioned in Chapter 1, in this book, we distinguish different kinds of ontologies: core, task and domain ontologies. A core ontology defines the generic model that can be further reused in task ontologies and specialized for different domains or fields of applications. A task ontology focuses on the actions, processes, and tasks that can be performed, often independent of a specific domain. A domain ontology focuses on the concepts and relationships specific to a particular area of knowledge denoting the domain.

Research question Q2.1 is addressed through a task ontology called PO² (Weber et al., 2023; Munch et al., 2022c; Ibanescu et al., 2016), described in Section 2.2.1. This ontology will cover the aspects related to the processing of inputs as well as the characterization of their properties during processing.

Standard languages exist to define ontologies, such as the Open Biological and Biomedical Ontology (OBO) format (Jackson et al., 2021) and the Web Ontology Language (OWL) (Hitzler et al., 2009) supported by the World Wide Web Consortium (W3C). In the OBO Foundry community, some ontologies deal with process modeling in biology: there are some generic ontologies, such as Information Artifact Ontology (IAO), Relation Ontology (RO), and Basic Formal Ontology (BFO) (Arp et al., 2015), as well as more specific ontologies, such as the Ontology for Biomedical Investigations (OBI) (Brinkman et al., 2010; Bandrowski et al., 2016) or the FoodOn Ontology (Dooley et al., 2018). Among W3C standards, the Semantic Sensor Network (SSN) ontology describes sensors and their observations, involved procedures, features of interest, samples used, and observed properties, as well as actuators. It includes a lightweight but

self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) (Janowicz et al., 2019). SSN and SOSA are the current recommendations promoted by the Open Geospatial Consortium (OGC) and the W3C to describe observational data. The Ontology for Food Process Experiment (OFPE) in Muljarto et al. (2014) proposes a generic semantic food transformation model based on OWL, which is built on four main classes: Product, Operation, Attribute, and Observation. OFPE is extended with the PO² ontology (Process and Observation Ontology), our proposal in this chapter (see Section 2.2.1). PO² allows us to represent a generic processing process described using a set of experimental observations available for the inputs and outputs for each step of the production process. A food process modeling working group compared different existing process models and provided a generalized process ontology in Dooley et al. (2024) and Dooley and Naravane (2023). Among the investigated models were OBI (Ontology for Biomedical Investigations) (Bandrowski et al., 2016), Exact2 (EXperimental ACTions) (Soldatova et al., 2014), Time Ontology in OWL²⁰, SOSA (Janowicz et al., 2019), PROV ontology²¹, and PO². The main problem in solving research question Q2.1 using OBI or Exact2 as a general process ontology is that they do not incorporate time, whereas PO² does by integrating the Time Ontology.

Research question Q2.2 is addressed with a task ontology, called CAPEX, described in Section 2.2.2. For a given situation of interest (food or biobased product/observation quality or defect), this ontology will provide all known possible explanations organized in a decision tree that starts from the most general explanatory situations, which must be refined with more specific explanations until precise enough to propose an action lever and an associated recommended technological action.

Fault tree analysis (FTA), which targets fault event risk assessment (Baig et al., 2013), can be compared to our approach. FTA enables computation of a level of risk represented by the occurrence probability of an undesired event. FTA also helps identify critical safety solutions to avoid the risk. For example, Pahasup-anan et al. (2022) analyzed different situations that could trigger a dust explosion in an extruded food production facility. Kim et al. (2020) used FTA to assess the level of risk of four situations that could help determine the risk of microbial contamination of food by *E. coli*. A fault tree includes a root node representing the undesired event. The tree's branches represent the explanatory scenarios, which may explain the fault event starting from basic events representing situations that would likely contribute to the overall fault defined in the roots. A whole fault tree could be considered as a set of scenarios associated with a probability of occurrence. Risk analysis is, of course, essential. However, FTA quantitative analysis requires collection of basic event occurrence measurements.

CAPEX, our task ontology (Buche et al., 2023), defines the decision tree structure, and is comparable to the tree structure used in FTA because we also represent a decision tree linking explanatory situations to a given situation of interest. However, our objectives are quite different in solving the research question Q2.2. Our original contribution compared to FTA consists in:

- proposing a semantic decision tree representation using Semantic Web languages to enable easier open data linkage with other sources of information available on the Web;

20. Time Ontology in OWL (2022, November 15; <https://www.w3.org/TR/owl-time/>).

21. PROV-O (2013, April 30; <https://www.w3.org/TR/prov-o/>).

- representing levers and associated technological actions to solve a situation of interest;
- representing the action efficiency based on individual experiences;
- representing contextual criteria associated with recommendations to help filter recommendations using a multicriteria approach; and
- identifying possible side effects associated with the implementation of a recommendation.

Moreover, FTA aims to estimate the probabilistic risk of failure, which requires the availability and collection of a substantial amount of numerical data, whereas our approach may be considered as more frugal as it is based on collecting and representing the collective technical know-how available for a given domain (company, food chain, etc.).

The CAPEX task ontology can be compared with the Core Ontology of Ghrab et al. (2016), known as the COOK ontology. In COOK, know-how is defined as the capacity/disposition to perform an action; this know-how concept is similar to our situation concept. COOK proposes a rich taxonomy to categorize different kinds of know-how (individual, collective, internal, external, crucial, etc.). From this perspective, we could consider that the situation concept in the CAPEX task ontology is a specialization of collective know-how. Our ontology enables us to represent complex interactions between situations (conjunction, reinforcement, weakening), which is not possible in COOK. In COOK, the knowing-that concept – a kind of belief – represents the relation between a proposition and a thinker. It assigns a truth value to the proposition. It is harder to compare this part of the COOK ontology to ours. Indeed, in our ontology, we implicitly consider that we represent a collective belief state, i.e., a COOK concept. However, we propose a more elaborate representation of the COOK proposition concept as we represent expert reasoning using the notion of possible causality between situations, the efficiency of an action, and associated contextual criteria. In short, there are several similarities between both ontologies with a richer description of kinds of know-how in COOK and an explicit representation of expert reasoning in our ontology, which COOK does not have. The part of our ontology that basically corresponds to the knowing-that concept in COOK is more detailed because we have proposed a complete and operational multicriteria decision support system (MCDSS) based on our ontology. Our ontology, like that of COOK, may be applied to any application domain based on know-how.

Compared to Buche et al. (2019), the preliminary research that led to the study described in this chapter, several new contributions are proposed and assessed:

- a model to aggregate action efficiency based on individual experiences;
- an extended version of the ontology expressed in OWL, including the representation of action efficiency and associated information (key and complementary criteria);
- and the definition and implementation of the action view.

Research question Q2.3 is addressed with a task ontology, naRyQ, described in Section 2.2.3. This ontology is designed for the data table integration (annotation and querying) task, also known as semantic table interpretation (STI) or semantic table annotation (Liu et al., 2023).

We published a method for table annotation using the task ontology naRyQ to solve research question Q2.3 (Touhami et al., 2011; Buche et al., 2013a, 2013b), which is

implemented in the @Web (Annotated Table of the Web) web application and its associated knowledge graph. Considering the annotation task typology proposed in Liu et al. (2023), our method performs Cell-Entity Annotation (CEA), Column-Type Annotation (CTA) and Row-to-Instance Annotation (RIA). It consists in annotating the cell (resp. column and row) using a concept from the task ontology. In this chapter, we assume that this annotation task is done manually (Lentschat et al., 2021a, 2022c; Buche et al., 2021b; Loustau-Cazalet et al., 2016; Guillard et al., 2016; Münch et al., 2023a). Automation of this task is discussed in Chapter 3.

The naRyQ task ontology (Buche et al., 2013a) was designed to represent experiments in which each experiment involves a studied object, some controlled or observed parameters, and a result measured for this object. In the field of food and biobased products, this generic model is well suited to represent characterization experiments in a simple tabular format (Guillard et al., 2016; Yun et al., 2018; Lentschat et al., 2022c; Münch et al., 2023a). However, even if it has been successfully used in simplified situations (Loustau-Cazalet et al., 2016; Fabre et al., 2020), this model is not well suited to represent any processing or characterization process involving a sequence of steps associated with several input and output components. This is why PO² (Weber et al., 2023) was designed to represent any process described by a set of processing/characterization steps and experimental observations including input and output components for each step of a process. The only tool comparable with @Web is, to the best of our knowledge, Rosanne (Rijgersberg et al., 2011), an Excel “add-in” application built upon the OM ontology (Rijgersberg et al., 2013), an ontology of quantities and units of measurement. Rosanne allows quantities and units of measurement associated with columns of an Excel table to be annotated using concepts from OM. As @Web does, Rosanne manages the notion of “phenomenon,” which is similar to subclasses from the symbolic concept notion in @Web, which represents nonnumerical data (e.g., step name or process name). The main difference is that @Web defines the notion of n-ary relation, which links together several pieces of data in the ontology (processing or characterization step associated with studied objects with controlled parameters and results) to represent a whole characterization/processing experiment.

Estimating data reliability (research question Q2.4) is crucial for scientists who want to reuse external data sources, since data are used to draw new conclusions. During data production, reliability is ensured through calibration, experimental design, and statistical repetition. However, traceability may be lost when data are collected from external resources such as scientific papers. When a validated physical model exists, data can be compared to the model’s predictions to assess reliability. In the absence of such models, other methods are needed, especially in fields with scarce data, such as the life sciences. To solve research question Q2.4, we present a method for assessing data reliability based on meta-information (Destercke et al., 2013). Multiple criteria are used, each contributing a different perspective on the data’s reliability. These individual insights are then combined into an overall evaluation, which is returned to the end user after appropriate post-processing. Our method offers the following key features:

- It handles conflicting information, since different criteria may yield contradictory assessments (e.g., data published in a reputable journal may have been gathered using unreliable instruments).

- It ensures traceability, allowing users to identify sources of conflict or, in their absence, understand the reasons behind a low or high reliability score.
- It is user-friendly and interpretable, both in terms of inputs and outputs, making it accessible to users who are not computer science experts.

The first three research questions presented in this chapter were addressed using the Web Ontology Language (OWL) standardized with the W3C. A presentation of this language and its associated terminology is given in Appendix 1 (more specifically, Appendix Sections 3, 4 and 5).

►► 2.2 Proposed solutions

In each of the following sections, we detail our solutions to the first three research questions of this chapter: first we present the competency questions (CQs) and the designed task ontologies. CQs represent functional requirements in the sense that the ontology and the developed ontology-based information system should be able to answer them (Uschold and Gruninger, 1996). Then, we focus on the implemented workflows designed to enrich and query the knowledge graphs associated with the ontologies. The solution proposed to solve research question Q2.4 is embedded in the workflow associated with the third research question Q2.3.

2.2.1 Solution proposed to represent generic knowledge and experimental data associated with a transformation process

This section deals with our solution to research question Q2.1 on how to represent a process of transforming raw materials into intermediate and final products and the associated observations.

We built the first version of the PO² task ontology using the basic NeOn scenario “from scratch” in 2016 (Ibanescu et al., 2016). In 2018, the Ontological Resources Reuse and Reengineering NeOn Scenario was used to align this model with existing ontologies (e.g., SOSA, BFO, IAO, and Time Ontology) for interoperability purposes (Munch et al., 2022c). The 2.3 version of PO² task ontology (Weber et al., 2023) included new properties (e.g., PO²:hasObjectOfInterest) and two classes from SOSA (sosa:Platform and sosa:System). The 2.4 version of PO² uses PROV-O, an ontology for representing the provenance information, adding metadata for documenting the agent who executed a step or an observation (Figure 2.1). The main competency questions (CQs) identified when building PO² are:

- CQ1: Which steps compose a given processing process?
- CQ2: Which attribute values are associated with controlled/observed parameters for each step?
- CQ3: Which are the observations where a specific attribute is observed for a given input (or output)?
- CQ4: What are the changes for an attribute value of an input during a given step?
- CQ5: What are all the steps and observations executed by a specific person or organization?

Figure 2.1 shows an excerpt of the PO² ontology (version 2.4), illustrating how PO² reuses classes and object/data type properties from several existing ontologies. SSN/SOSA classes are shown in green, the Time Ontology class TemporalEntity in light blue, PROV-O classes in yellow, and PO² classes in dark blue.

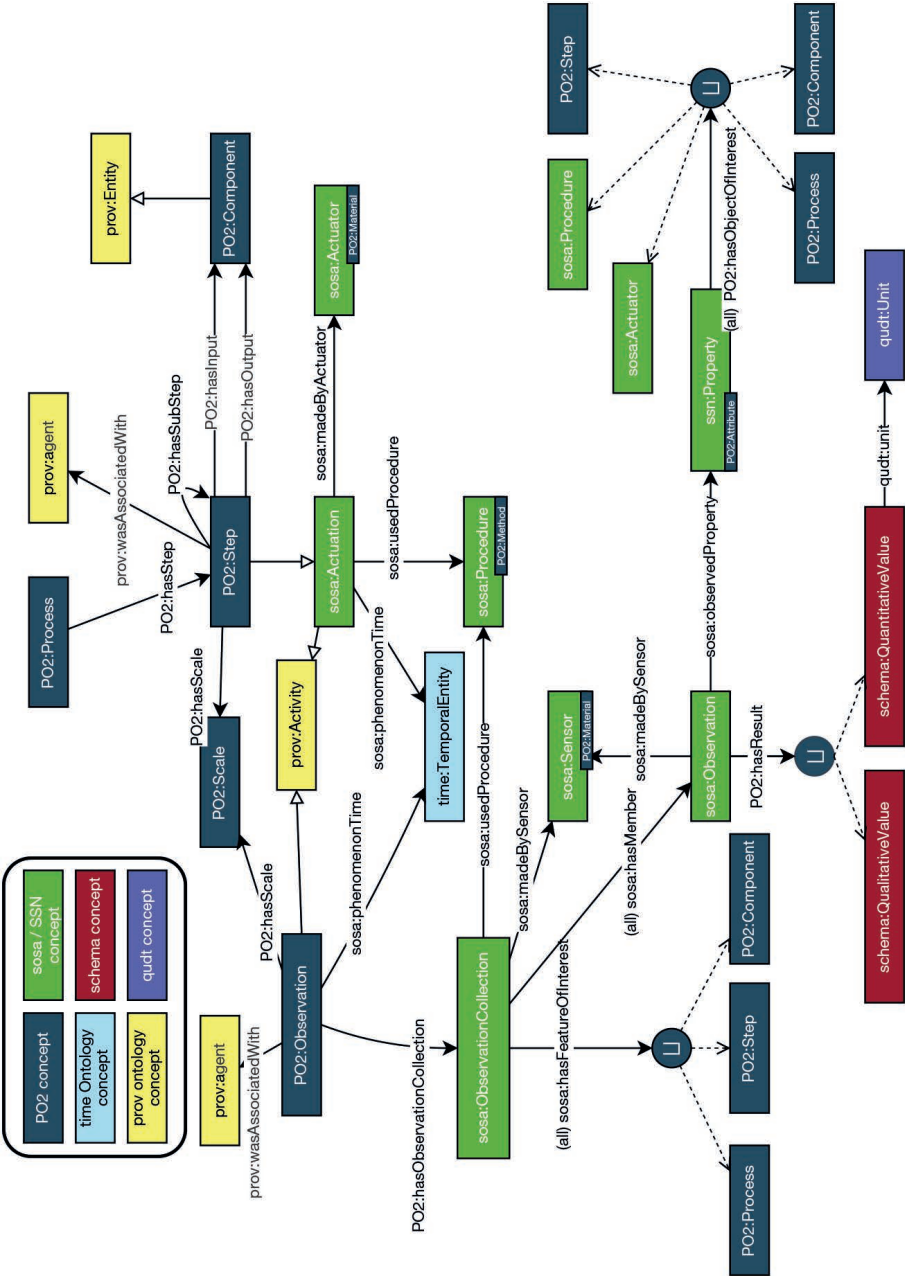


Figure 2.1. Excerpt of 2.4 version of PO² ontology (drawn using Chowik graphical language).

The main classes of PO² and their ObjectProperties are:

- The PO²:Process class and PO²:Step class are linked by the ObjectProperty PO²:hasStep; this makes it possible to represent the steps composing a transformation process and is the answer to CQ1: Which steps compose a given processing process?

- The PO²:Step class represents a unit operation that transforms inputs into outputs; inputs and outputs are represented as a PO²:Component class and are linked using the ObjectProperties PO²:hasInput and PO²:hasOutput, respectively. Each PO²:Step is a subclass of the *sosa:Actuation* class, which is characterized by a temporal duration. It is performed by a device, represented by the *sosa:Actuator* class, using a given method represented by the *sosa:Procedure* class. This is the modeling answer to CQ2: Which attribute values are associated with controlled/observed parameters for each step?

- The PO²:Observation class represents a collection of *sosa:Observation* classes and is characterized by a temporal duration; the ObjectProperty PO²:hasObservation Collection links PO²:Observation class and *sosa:ObservationCollection* class. All members of the *sosa:ObservationCollection* are performed by a sensor, represented by a *sosa:Sensor* class. The *sosa:Observation* class is linked to the *ssn:Property* class with the *sosa:observedProperty* relation, in order to associate an observable property. A *sosa:ObservationCollection* is linked with an input (or output) by the *sosa:Feature-OfInterest* relation having as range the union of three classes: PO²:Process, PO²:Step, and PO²:Component. This is the modeling answer to CQ3: Which are the observations where a specific attribute is observed for a given input (or output)?

- Using the *sosa:ObservationCollection* class and *sosa:Observation* class linked by the *sosa:hasMember* relation makes it possible to represent time-series data; this is the answer to CQ4: What are the changes for an attribute value of an input during a given step? The values obtained for the observed properties are linked to *sosa:Observations* thanks to the PO²:hasResult data type property.

- The PO²:Step class and the PO²:Observation class are linked with the *prov:agent* class, making it possible to find/track the agent. This is the modeling answer to CQ5: What are all the steps and observations executed by a specific person or organization?

In our PO² ecosystem, the *ssn:Property* class is labeled PO²:Attribute, the *sosa:Actuator* class is labeled PO²:Material, and the *sosa:Procedure* class is labeled PO²:Method. In addition, PO²:Step, PO²:Process and PO²:Observation are subClassOf *prov:Activity* and PO²Component subClassOf *prov:Entity*.

When building a domain ontology that extends this PO² task ontology, the subclasses of the following main classes must be defined: PO²:Component, PO²:Step, PO²:Attribute (*ssn:Property*), PO²:Observation, PO²:Material (*sosa:Actuator*), and PO²:Method (*sosa:Procedure*). The method to build the TransformON domain ontology is presented in Chapter 1, Section 1.2.1, with a detailed description of the classes in Section 1.3.1.

Figure 2.2(a) shows the workflow designed to enrich TransformON domain ontology with the RDF data part: the PO²Manager tool has been presented in Chapter 1, Section 1.2.1. Figure 2.2(b) shows the workflow designed to query the PO² knowledge database using an application called SPO²Q (Simple PO² Query). We will now present a more detailed example that illustrates functionalities of PO² ecosystem (PO²Manager and SPO²Q) to facilitate experiment reproducibility.

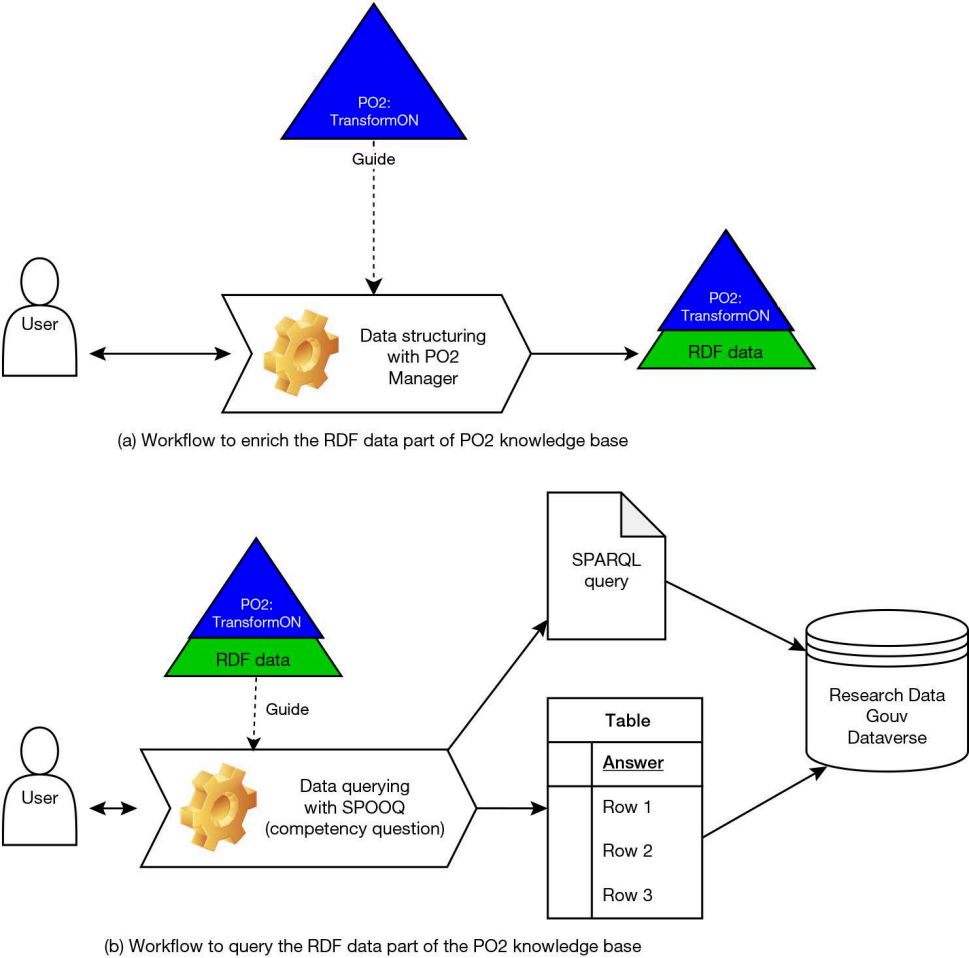


Figure 2.2. Workflows associated with the PO² knowledge graph.

An example of a vegetal-powder production itinerary (Itinerary 3), excerpted from Fabre et al. (2020), is shown in Figure 2.3. The red nodes are the input and output of a step; the black and green nodes are steps with observation(s); the black arrow between steps indicates the temporal order of the steps; the red arrows indicate the input/output of a step. The itinerary represents a sequence of two milling steps: the cutting milling step followed by the impact milling of the input (i.e., the hemp core (1) (CQ1)). Controlled parameters associated with the second milling step performed using the UPZ impact mill are presented in Figure 2.4: the rotation speed is 18,000 rpm/min, the fill rate is 1 kg/h, and the sieving size is 0.3 mm. Storing these values will allow the same experiment in the future (CQ2). FAIR principles are also implemented when storing observations. For example, Figure 2.5 shows the granulometric observation that was made using the Malvern 2000 equipment when using standard operating procedure 1 as described in Fabre et al. (2020). The recorded value for raw data on line 0 is that the median particle size is between 143 and 163 micrometers (CQ3).

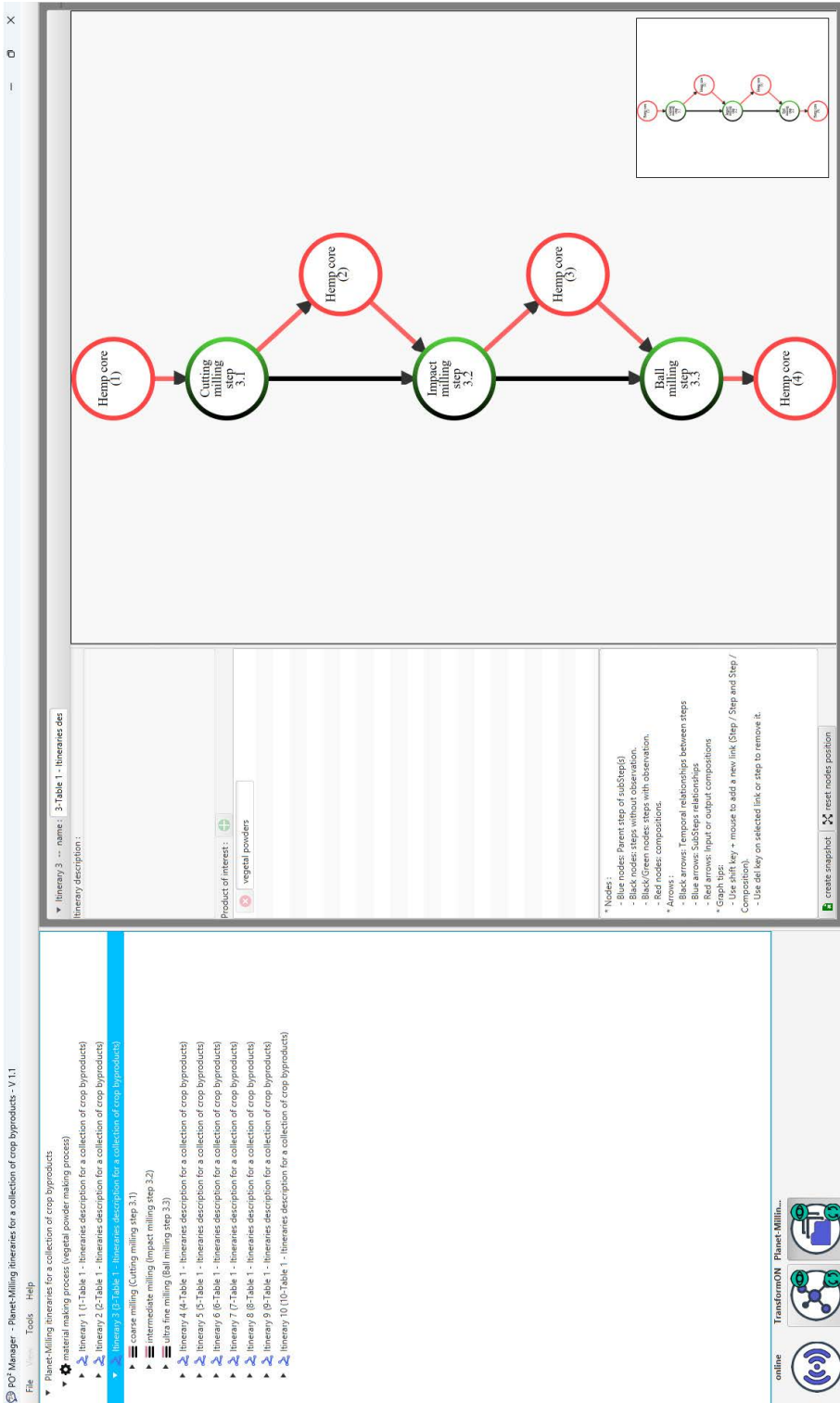


Figure 2.3. PO² Manager screenshot of an excerpt of an example itinerary to produce a vegetal powder.

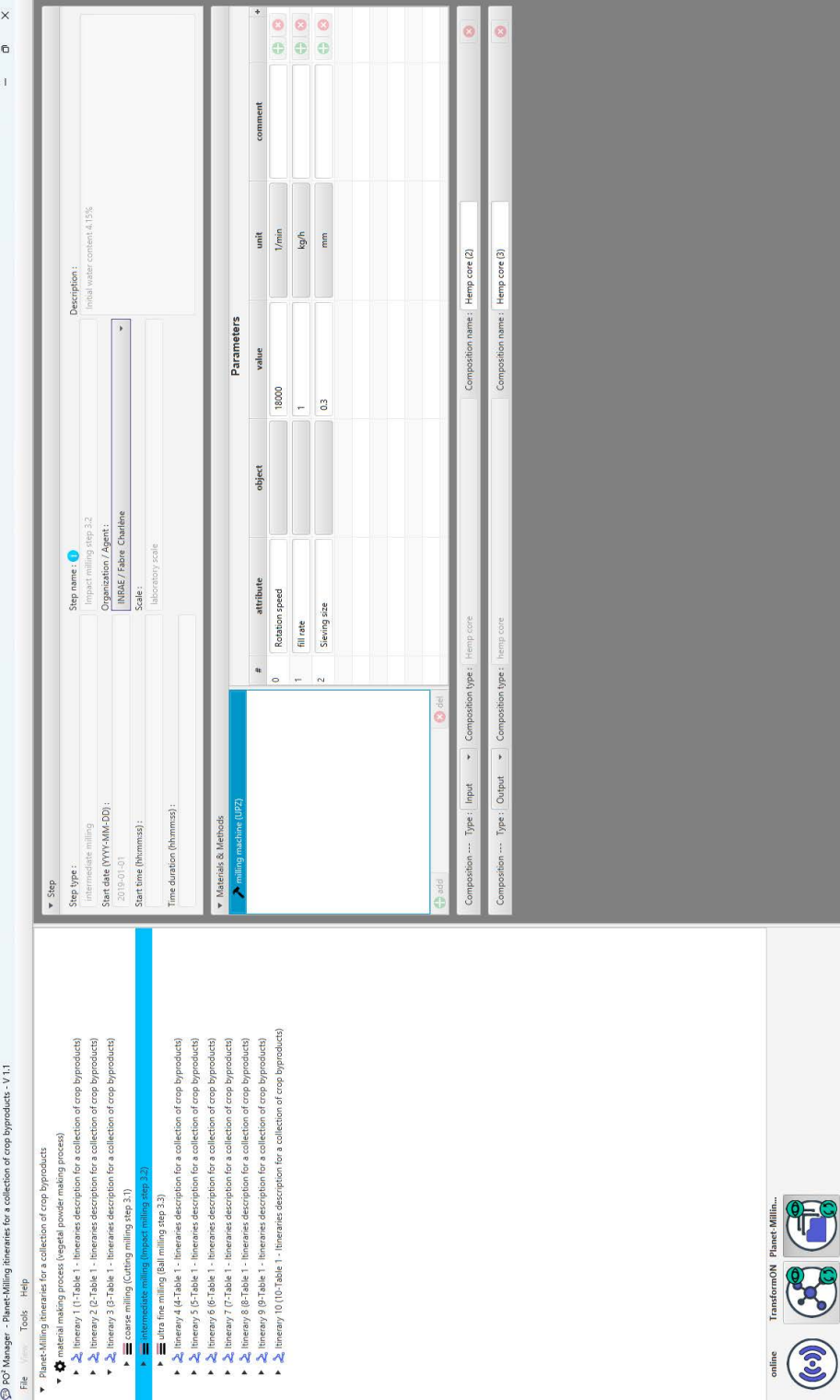


Figure 2.4. PO² Manager screenshot of an excerpt of controlled parameters associated with the impact milling shown in Figure 2.3.



Figure 2.5. PO3 Manager screenshot of an excerpt of granulometric observation associated with the impact milling shown in Figure 2.3.

Several years after Fabre et al. (2020) was published, the same team wanted to reproduce the same experiment (Itinerary 3) using the same input to assess reproducibility and produce some additional observations (e.g., the biochemical composition of the vegetal powder). First, the following query needed to be executed: “Which attribute values are associated with controlled/observed parameters (i.e., the volume D50) of each step?” in order to retrieve experiments performed on the input hemp core (1) that produced powders ranging in size between 140 and 160 micrometers. The query generated using the SPO²Q tool is shown in Figure 2.6. The results of this query are given in Figure 2.7 and retrieve the itinerary presented in Figure 2.3 and 2.4. The team could then easily reproduce the same experiment and published the results in Mayer-Laigle et al. (2024), showing that they could obtain more or less the same granulometric values (i.e., 160.6), as shown in Figure 2.8, for Itinerary 5.

Figure 2.2(b) shows the workflow designed to query the PO² knowledge database. The identified CQs were implemented through the SPO²Q web application. This application was developed specifically for scientists, who are not necessarily computer scientists, to lower the barrier to querying complex knowledge graphs. The idea is to provide a guided form-based interface so that users can query the data without needing to know either SPARQL or the internal schema of the knowledge graph. For advanced users, an expert mode is also available, allowing them to edit or refine the generated SPARQL queries when the constraints of the form are not sufficient.

SPO²Q relies on the structure of the PO² task ontology to guide the user through the construction of a query. The interface is organized into successive blocks, corresponding to the main PO² classes: process, step, and observation. The chaining of these blocks is ontology-driven: once a process is selected in the first block, the application automatically checks the dataset to identify which steps are instantiated for this process, and only these steps are proposed in the next block. Similarly, when a step is selected, the possible observations associated with it are retrieved and suggested to the user.

The vocabulary available in the SPO²Q form fields is therefore directly constrained by the domain ontologies used to annotate the dataset. Domain ontologies such as TransformON are embedded in the datasets themselves: when a dataset is selected, the corresponding ontology is used to restrict the values available for query construction. This ensures both the semantic consistency of the queries and the reusability of the controlled vocabulary across datasets.

Once a query is built and executed against the RDF knowledge graph, SPO²Q provides different format output options. The results can be downloaded in TSV or JSON formats, making them directly usable for further statistical analyses or integration into external tools. Queries can also be saved for later reuse, facilitating reproducibility and collaborative work among scientists.

An example is shown in Figure 2.6, where hemp core is selected as an input component and the values for volume D50 are fixed. This illustrates how SPO²Q allows users to formulate meaningful domain queries without having to manipulate RDF data directly. The RDFS reasoner is executed to automatically enrich the knowledge graph before query execution. Altogether, SPO²Q represents our software solution for user-friendly, ontology-driven querying of PO²-based knowledge graphs.

INRAE

Simple PO² Query

Form

← → ↺ ⌂

quantum.mis-ps.inrae.fr/spoq/form

Selected dataset

• PlanetMilling_literaries_for_a_collection_of_crop_byproducts

Selected process

by process name : doNotFilter

Step type : Step

Step Material

Input component

hemp core

Ingredient

Output component

Observation name : doNotFilter

Observation Material

volume D50

☐ Show objects

Filter type

≥

Value

140

Unit

um

Object Of Interest

volume D50

☐ Show objects

Filter type

≤

Value

160

Unit

um

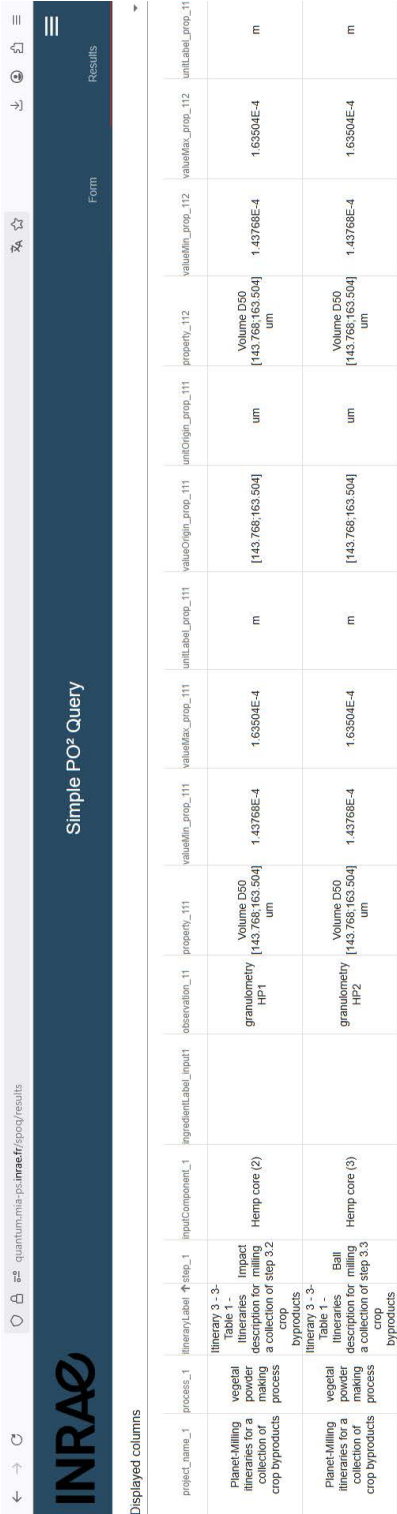
Object Of Interest

Add an observation

Add a step

Execute form

Figure 2.6. Querying the PO² knowledge graph with SPO²Q: focus on query generation.



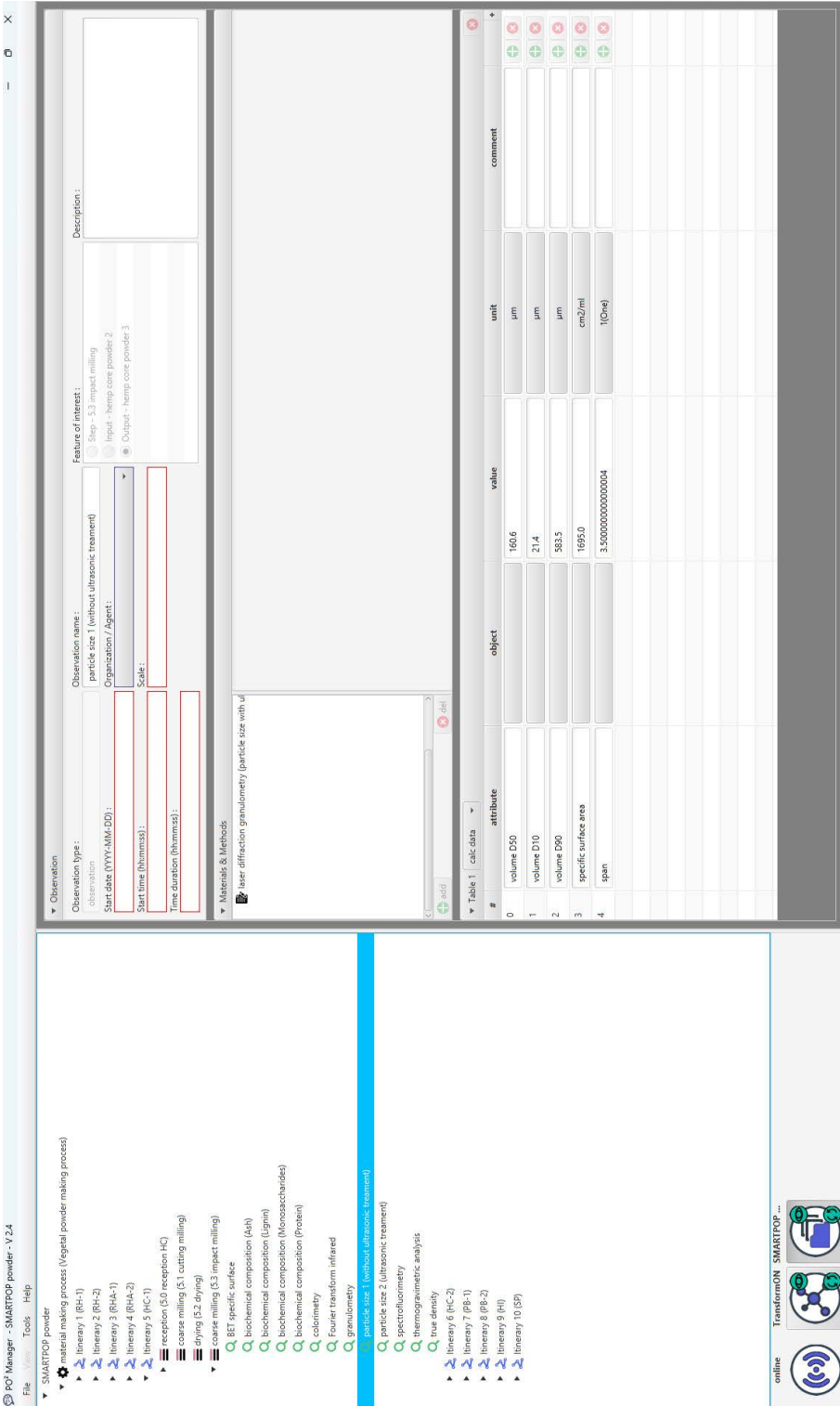


Figure 2.8. PO² Manager screenshot of an excerpt of granulometric values (i.e., 160.6 for volume D50) published in Mayer-Laigle et al. (2024) obtained after reproduction of the experiment presented in Figure 2.5.

Figure 2.9 illustrates a subset of RDF triplets from a three-step process workflow structured by the PO² ontology. On the right-hand side, an observation line (Volume D50) is visible for a specific step (Impact milling step 3.2), with triplets representing the quantitative values and the unit of measurement used. It corresponds to the Volume D50 attribute value shown in Figure 2.5 using PO² Manager.

To date, three data papers (Munch et al., 2022b, 2025; Mayer-Laigle et al., 2024) have been published to describe data stored in the PO² knowledge graph. More details are provided in the 2.3 Outcomes section below.

2.2.2 Solution proposed to represent expert and operational collective know-how

In this section, we lay out our solution to answer the research question Q2.2 on how to represent expert and operational know-how. Typical CQs and requirements addressed by CAPEX are:

CQ1: For a given situation of interest (targeted quality or default to tackle), what are all the possible explanations organized in a decision tree starting from the more general explanatory situations that must be refined by more specific ones until it is precise enough to propose an action lever and an associated recommended technological action?

Additional requirements:

Interactions between explanatory situations must be represented. Two kinds of interactions must be considered: (i) conjunctive interactions of situations S1 and S2 to explain S3, meaning that situation S3 may appear only if S1 and S2 appear; (ii) reinforcement (resp. weakening) interaction of situation S1 by situation S2 to explain S3, meaning that the effect of S1 on S3 is reinforced (resp. weakened) if S2 appears.

Explanation: The decision tree will permit a user to consider all possible known explanations of a given situation of interest. This functionality, mostly dedicated to junior technicians, is very important in food production chains, which must deal with high staff turnover.

CQ2: For a given situation (situation “of interest” or explanatory), what is the value of a relevant analytical parameter that allows the user to verify that this situation is really happening?

Explanation: This is of considerable interest for both young and experienced technicians who must deal with several specific production processes (e.g., the *Excessive salting* situation of interest in Comté is associated with the analytical parameter NaCl > XXX g/100 g cheese) and have not memorized all the analytical parameter values associated with encountered situations.

CQ3: What is the feedback of technicians’ individual experiences in terms of technological action efficiency to treat a given situation of interest?

Explanation: Action ranking based on efficiency is extremely important to help users choose the “best” action to solve a given situation of interest.

CQ4: What are the contextual criteria relevant for decision support and associated with those assessments?

Explanation: It is required to avoid any ambiguity or misunderstanding of propositions made by the multicriteria decision support system (MCDSS). For example, a given action may be considered as very efficient in the short term (ST), and not at all efficient in the long term (LT). In this example, the criterion ST/LT must be defined as a key criterion because it determines the action efficiency. Moreover, additional criteria may be associated to facilitate decision support depending on the chain (economical cost, sanitary risk, etc.). A detailed presentation of CAPEX MCDSS is presented in Chapter 4. This section is dedicated to the presentation of the CAPEX knowledge graph creation.

Figure 2.10 is a graphical representation of the CAPEX OWL ontology. As expressed in CQ1, a situation S1 (an instance of the *Situation* OWL class) is explained by a situation S2 through an instance of the *CausalityNode* class linked to S1 (resp. S2) by the OWL *hasForCause* (resp. *hasForConsequence*) object property. Note that S1 may be an instance of the *SituationOfInterest* class that is a kind of situation. A situation S may be associated with an action A via the *hasForAction* object property. An action A is associated with its lever through the *hasForLever* object property. A conjunctive interaction CI1 (an instance of the *SituationConjunction* class) is linked to conjunctive causal situations S1 and S2 (and other situations if required) by the *isComposedOf* property. CI1 is linked to an instance of the *CausalityNode* class by the *hasForCause* property. This *CausalityNode* class instance is linked to the consequence situation S3 by the *hasForConsequence* property. The strengthening (resp. weakening) interaction of S1 by S2 to explain S3 is also represented using a conjunctive interaction CI1 (an instance of the *SituationConjunction* class). The asymmetric role of situations is achieved as follows: the altered situation S1 is linked to altering situation S2 via the *SpecificationOfWeakening* (resp. *SpecificationOfReinforcement*) object property if the alteration type is weakening (resp. strengthening).

The *isDetectedBy* datatype property associated with an instance of the *Situation* class implements CQ2. An instance of *Action* is associated with an instance of the *Efficiency* class to implement CQ3. The *hasForKeyCriterion* datatype property is used to determine the list of criteria values associated with a single *Efficiency* instance (CQ4). The *hasForScore*, *hasForObservations*, and *hasForTechniciansAgreement* datatype properties are associated with an *Efficiency* class instance that is linked to an *Action* instance. The *hasForConsequenceCriterion* object property links an *Efficiency* instance with a set of pairs (name, value) that are used for decision support. The *refersToDefect* object property links an *Efficiency* instance with the situation of interest to which it refers.

Buche et al. (2019) proposed a method that enables collective mind mapping dedicated to the CAPEX knowledge graph enrichment. Interested readers may refer to that paper for further details on this topic. Any mind-mapping tool may be used to edit decision trees. Interested readers should check that CAPEX knowledge graph enrichment succeeds using imports of a markdown file including the decision tree generated by the mind-mapping tool.

Figure 2.11 presents the two workflows of the CAPEX system presented in Buche et al. (2022). Figure 2.11(a) shows the subsystem that automatically generates the CAPEX RDF knowledge graph from a mind-mapping markdown file representing

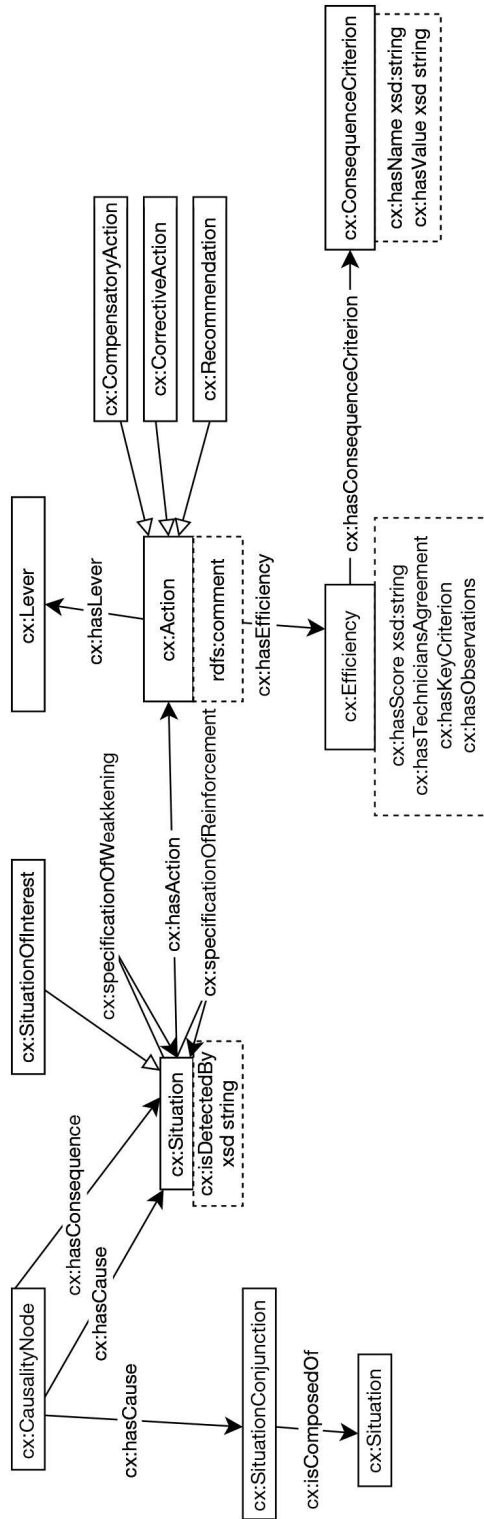
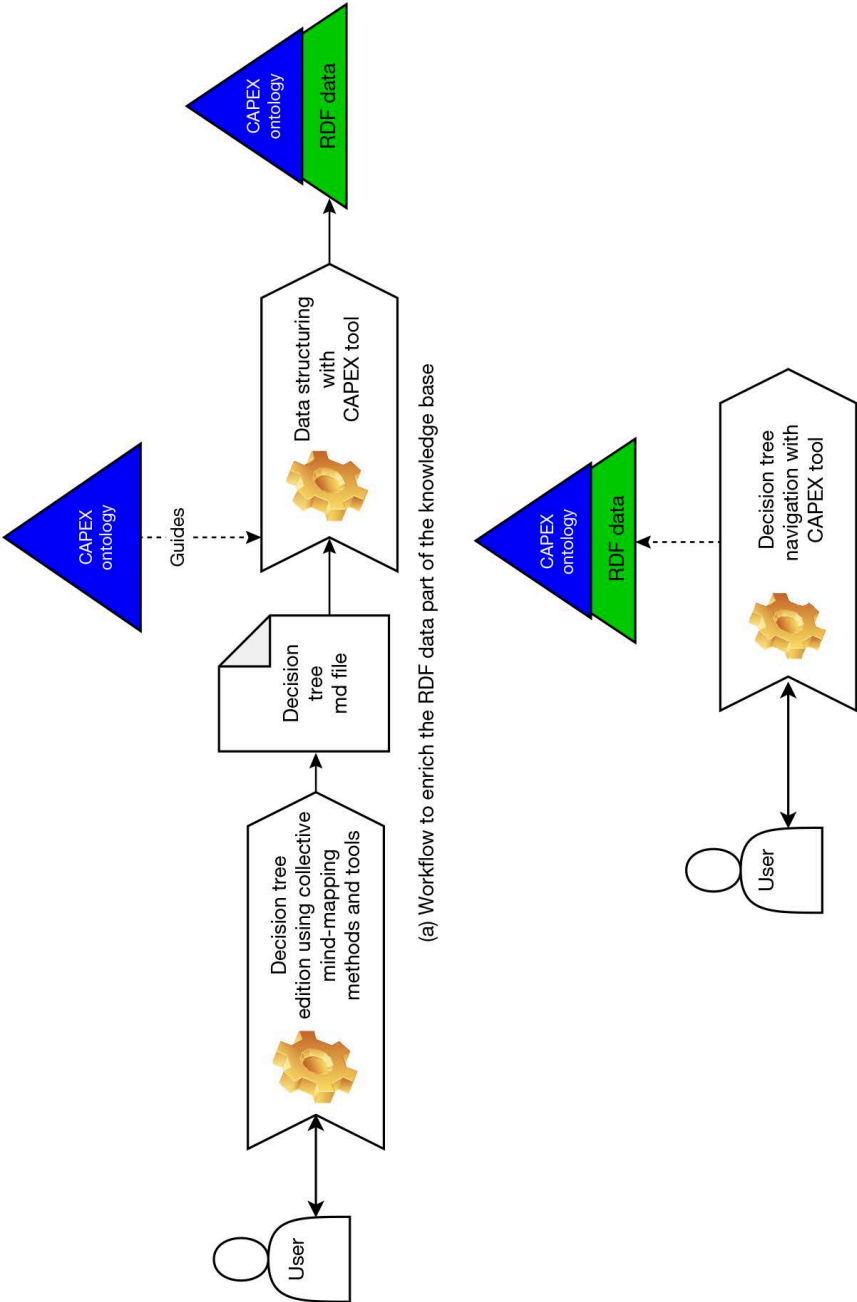


Figure 2.10. CAPEX ontology inspired by Buche et al. (2023), drawn using Chowik graphical language.



(b) Workflow to query the RDF data part of the knowledge base

Figure 2.11. Workflows associated with the CAPEX knowledge graph.

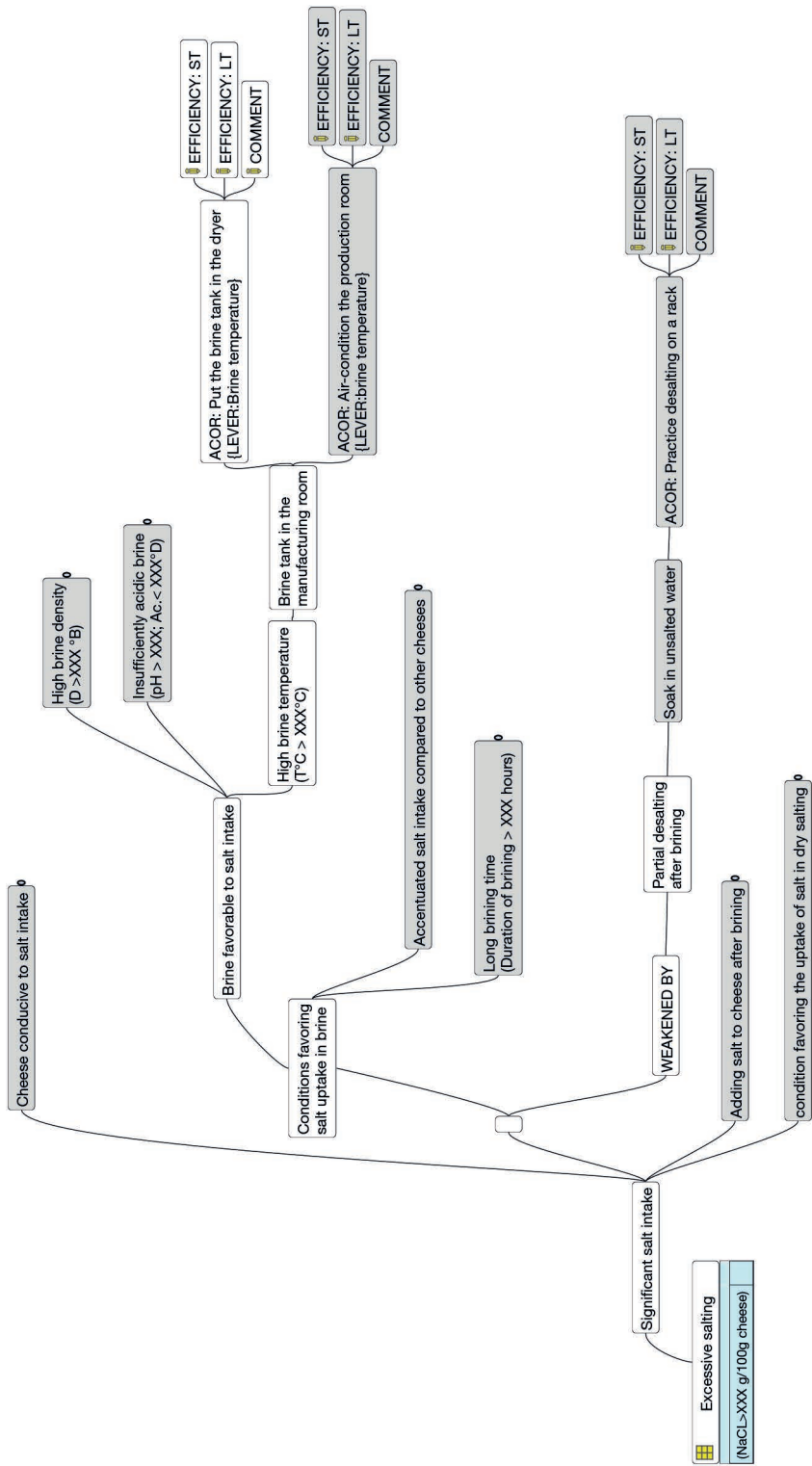


Figure 2.12. An excerpt of the mind-mapping file associated with the *Excessive salting* situation of interest reproduced from Buche et al. (2023).

the decision tree associated with the situation of interest. Figure 2.11(b) presents the CAPEX RDF knowledge graph querying subsystem that can be used to deliver decision support. The RDFS reasoner is executed to automatically enrich the RDF knowledge graph before query execution. This part of the system is described in more detail in Section 4.2.1. Figure 2.12 is an excerpt from a mind-mapping file representing the decision tree associated with the situation of interest: *Excessive salting*, depicted by the blue node on the bottom left side of the figure. The complete mind-mapping file is available in Buche et al. (2022). This *Excessive salting* situation of interest may be explained by the *Significant salt intake* situation. From there, four possible explanations emerge. We will focus on the one represented by a white node, namely *Conditions favoring salt uptake in brine*, along with its associated branch presenting three possible explanations. We will focus again on one of them, *Brine favorable to salt intake*, until reaching the two recommended corrective actions: *Put the brine tank in the dryer* and *Air-condition the production room*. Figure 2.13 provides a close-up of the table associated with the *EFFICIENCY: ST* node. This table includes the efficiency indicator along with the number of observations and the contextual criteria associated with them (see *hasForScore*, *hasForObservations*, *hasForTechniciansAgreement*, and *hasForConsequenceCriterion* properties presented above).

In Figure 2.14, we present part of the RDF knowledge graph corresponding to the translation of the branch whose nodes are white in the decision tree shown in Figure 2.12. For more information about CAPEX RDF knowledge graph querying for multicriteria decision support, please refer to Section 4.2.1.

To date, no data paper has been published to describe data stored in the CAPEX knowledge graph. However, two knowledge graphs have been created (see Section 2.3: Outcomes for more details).

2.2.3 Solution proposed to represent data tables associated with external sources

As mentioned in Section 2.1, the naRyQ task ontology (Buche et al., 2013a) was designed to represent experiments in which each experiment involves a studied object, some controlled or observed parameters, and a result measured on this object. In the field of food and biobased products, this generic model is well suited to represent characterization experiments in a simple tabular format such as what scientists typically use to register them. Figure 2.15 contains Table 4 from Vuillet et al. (2026): column 1 indicates the studied object (i.e., different biobased films used for packaging); columns 3-5 indicate measurements of H₂O permeability, which is the result of the experiment; column 2 indicates the controlled temperature of the measurement.

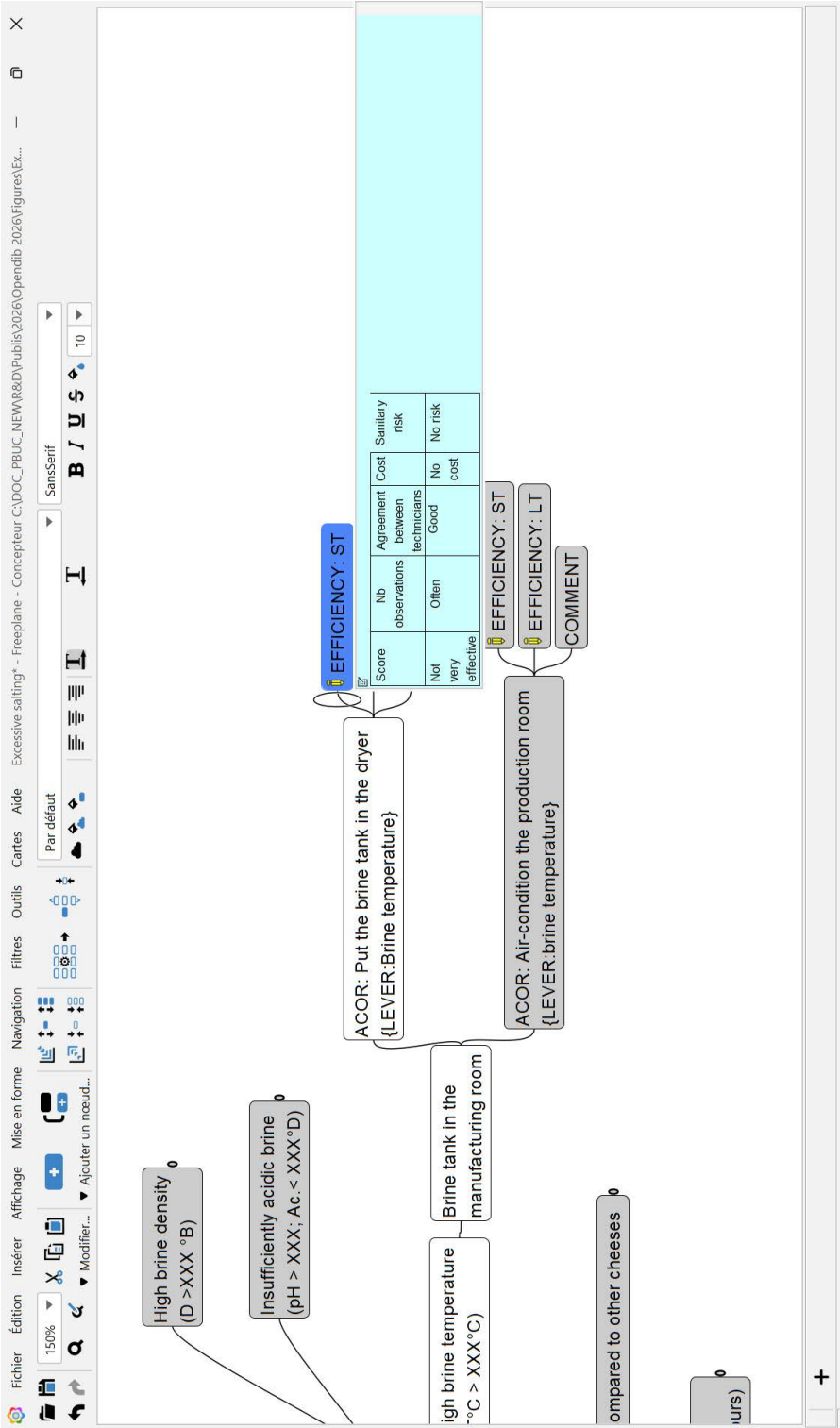


Figure 2.13. Close-up of the table associated with the *EFFICIENCY: ST* node associated with the *Put the brine tank in the dryer* node reproduced from Buche et al. (2023). ST: short term; LT: long term.

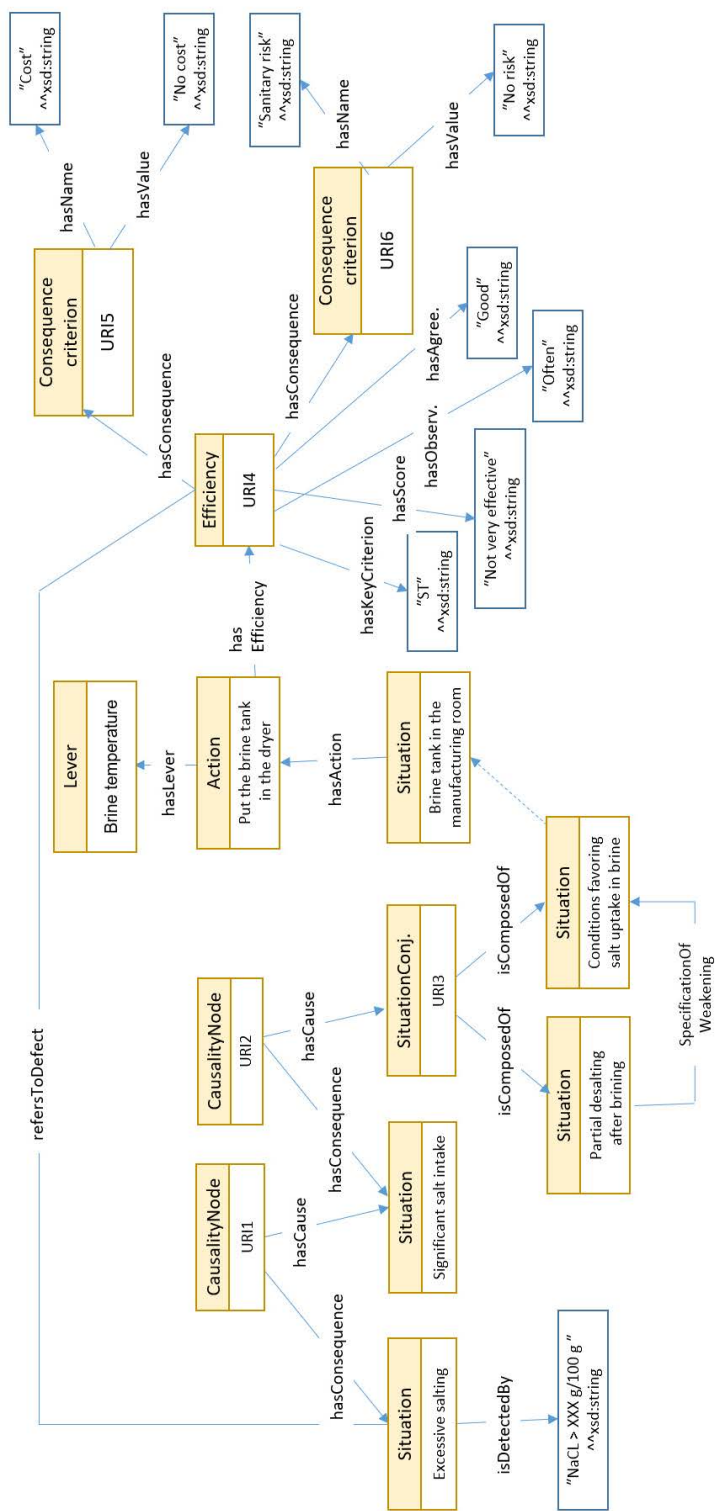


Figure 2.14. An excerpt of the CAPEX knowledge graph corresponding to the selected white branch in Figure 2.12 reproduced from Buche et al. (2023).

Table 4
 Water vapor permeability of CMC-based self-supported films as a function of water activity, a 10 % uncertainty can be considered.

	<i>T</i> (°C)	WVP (barrer)		
		<i>a_w</i>		
		0.5	0.6	0.7
CMC	25	44 200	52 100	74 800
	40	34 000	N/A	N/A
CMC-But- <i>TertB</i> _{0.13}	25	2500	3600	6100
	40	2800	4900	10 900
CMC-But- <i>TertB</i> _{0.30}	25	14 300	20 100	36 300
	40	8500	11 500	25 900
CMC-Norb- <i>TertB</i> _{0.35}	25	6800	8200	22 500
	40	7500	10 200	22 200

Figure 2.15. An excerpt of a data table published in a scientific paper (from Vuillet et al., 2026).

The main competency questions (CQs identified when building naRyQ) are:

- CQ1: Which studied object (e.g., food packaging) is characterized by a given experimental measurement (e.g., H₂O permeability)?
- CQ2: Which controlled/observed parameters (e.g., temperature or packaging thickness) are associated with attribute values (e.g., O₂ permeability, CO₂ or H₂O permeability)?
- CQ3: Which units of measurement are associated with a given experimental measurement (e.g., H₂O permeability)?
- CQ4: Which methods are associated with a given experimental measurement (e.g., H₂O permeability).

Figure 2.16 gives an excerpt of naRyQ, called an Ontological and Terminological Resource (OTR) in this chapter, in the matter transfer domain: this task ontology is called TRANSMAT (for MATter TRANSfer; see Guillard et al., 2018).

The main classes in the core conceptual component of naRyQ are:

- Relation_concept class, which is used to identify the result of the experiment.
- Argument class, which is used for the controlled/observed parameters; its subclasses are Symbolic_concept class and Quantity class.
- om:Unit class, which along with the Quantity class, allow the user to manage quantities and their associated units of measurement. This modeling choice was inspired in 2013 by the OM ontology (Rijgersberg et al., 2013), an ontology of quantities and units of measurement, as it seems to be well suited for representing all the units of measurement identified in the domain of food packaging.

The domain conceptual component of naRyQ is composed of a set of classes that specialize the one from the core conceptual component for a given application domain. In Figure 2.16, the Packaging class is a subclass of Symbolic_concept in the matter transfer domain. naRyQ has been designed to semi-automatically annotate

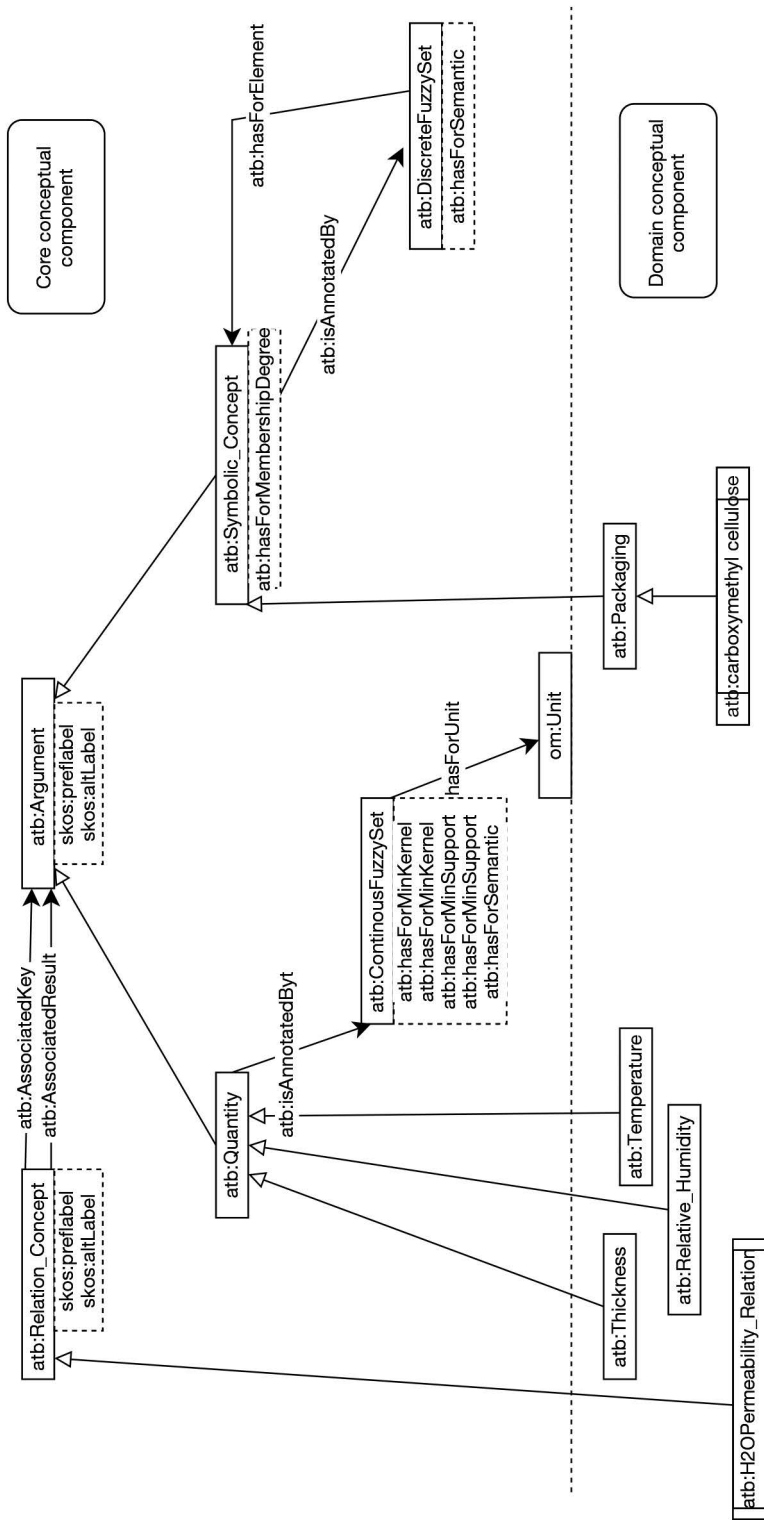


Figure 2.16. An excerpt of naRyQ in the TRANSMAT (matter transfer) domain inspired from Lentschat et al. (2022a) and drawn using Chow/k graphical language.

data tables from scientific papers, such as the one presented in Figure 2.15. A complete method to perform this text mining task is presented in Chapter 3. To be able to represent textual similarities in this text mining task, fuzzy sets were introduced (see the `DiscreteFuzzySet` class in Figure 2.16). To represent imprecise data (e.g., min–max values, mean and standard deviations) expressed in data tables, continuous fuzzy sets were introduced in `naRyQ` (see Figure 2.16). For more details about `naRyQ`, and more specifically the management of fuzzy values, see Buche et al. (2013a).

As shown in Figure 2.16, the terminological component of `naRyQ` contains the set of terms describing the studied domain that are used to annotate experimental data. Subclasses of the generic classes `Relation Concept`, `Symbolic Concept`, and `Quantity`, as well as instances of the generic class `Unit Concept`, are all denoted by at least one term from the terminological component. Each of those subclasses or instances are, in a given language, denoted by a preferred label and optionally by a set of alternative labels, which correspond to synonyms or abbreviations. Labels are associated with a class or an instance thanks to SKOS (Simple Knowledge Organization Scheme) labeling properties recommended by W3C²².

The `@Web` annotation module implements a complete workflow to extract either manually or semi-automatically experimental data tables from scientific documents (e.g., the table from Figure 2.15), and semantically annotate them with classes from `naRyQ`.

The four workflows of the `@Web` module are presented in Figure 2.17. In workflow (a), new classes belonging to a given domain ontology may be added in the `naRyQ` ontology to be used to annotate data tables retrieved from scientific documents. This will enrich the domain conceptual component presented in Figure 2.16. New domain classes are given in a predefined textual format, named “Excel template,” designed in Excel so as to be user-friendly for interaction with domain experts, and fully compliant with the `naRyQ` definition. The importation module of `@Web` reads these Excel files and automatically generates the updated `naRyQ` ontology in OWL format. Workflow (b) generates an Excel template for data entry, compliant with the set of `n`-ary relation classes already defined in the `naRyQ` ontology. After completing these files with data from data tables, the `@Web` data importation module will populate the `naRyQ` ontology and store new RDF data in the `@Web` knowledge graph during step (c). Figure 2.18 shows an example of metadata associated with a scientific document registered in the `@Web` knowledge graph. The example of its only annotated data table is shown in Figure 2.19: the original table from Figure 2.15 is annotated with the `H2OPermeability_relation` class from Figure 2.16. Optionally, in step (c), each document may be associated with a reliability assessment score (see Section 2.2.4 for more details). The annotated data tables are stored in an RDF knowledge graph (see Figure 2.20). The RDFS reasoner is used to automatically enrich the knowledge graph before query execution using the `@Web` querying subsystem in step (d). Readers interested by the `@Web` querying subsystem should read Lousteau-Cazalet et al. (2016) for more details.

Chapter 3 presents the semi-automatic method designed to annotate data tables using the `naRyQ` ontology.

To date, seven data papers have been published to describe data stored in the `@Web` knowledge graph (see Section 2.3.3 for more details).

22. Simple Knowledge Organization System Reference (2009, August 18; <https://www.w3.org/TR/skos-reference/>).

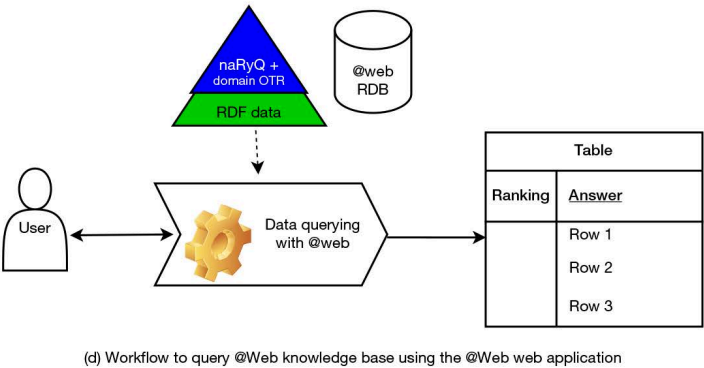
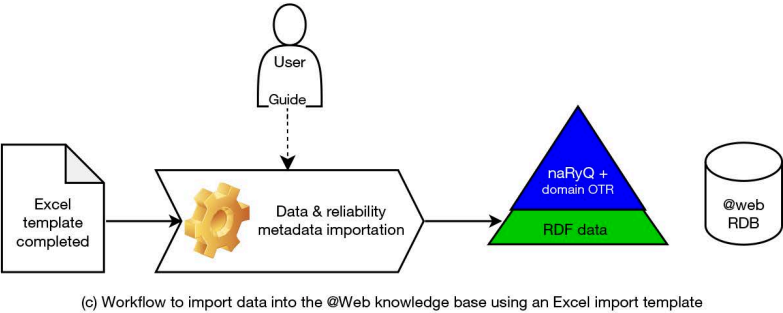
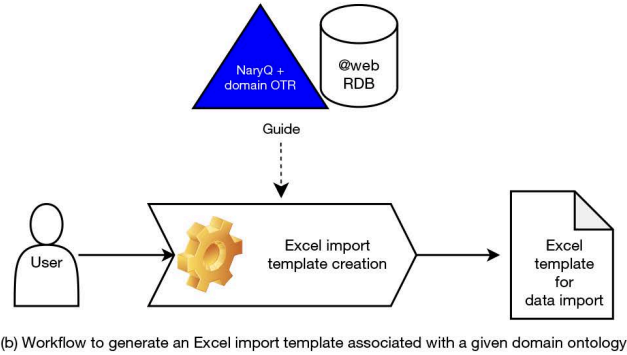
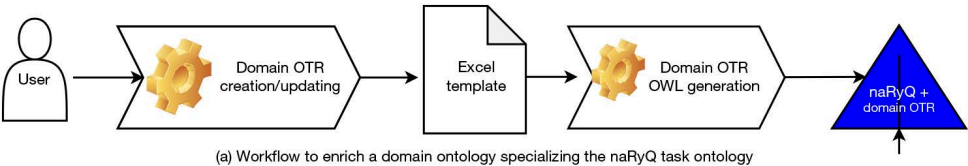


Figure 2.17. Workflows associated with the @Web knowledge graph.



Figure 2.18. Screenshot of @Web showing the metadata associated with a scientific document stored in the @Web knowledge graph.

Figure 2.19. Screenshot of @Web showing an example of an annotated data table with the permeability relation class from the naRyQ ontology.

Figure 2.19. Screenshot of @Web showing an example of an annotated data table with the permeability relation class from the naRyQ ontology.

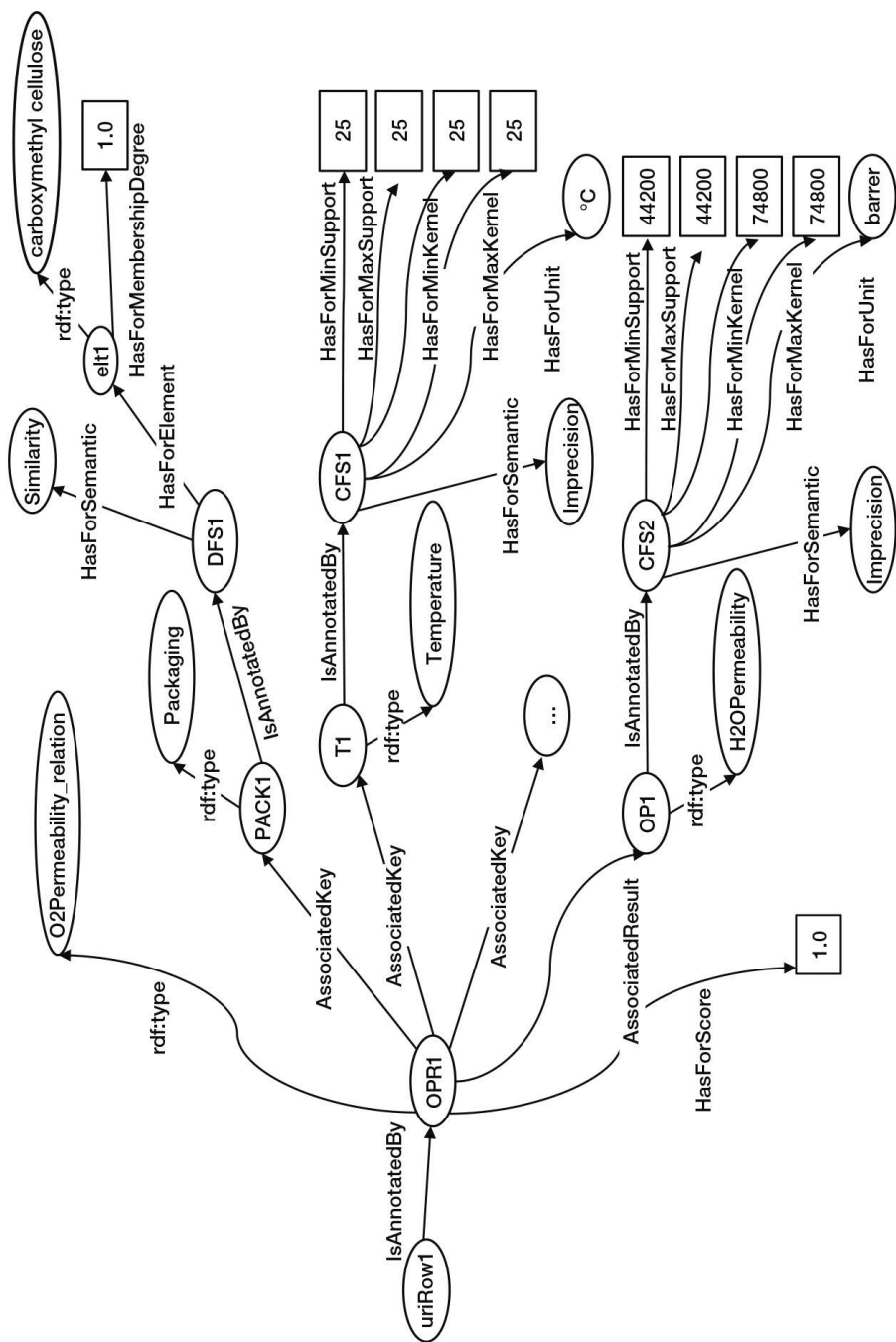


Figure 2.20. An excerpt of the @Web knowledge graph corresponding to the first line of the @Web screenshot shown in Figure 2.19.

2.2.4 Solution proposed to assess reliability of external sources

To assess the reliability of external data sources, an extra step was added to the @Web software workflow presented in Figure 2.17. This step was added between steps a and b to define the criteria used to assess reliability. These criteria are elements that are usually found in external sources as publications (reports, papers, etc.) reporting experimental results. Note that most of these criteria are not specific to the life sciences and can be used for any experimental data. The list of criteria below is, of course, not exhaustive:

- The first group concerns the data source itself. It contains features such as the source type (e.g., scientific publication, government report, web page), the source reputation (e.g., the category of journal where the article has been published), the number of times the source has been cited or the publication date (data freshness is important in the life sciences, due to the rapid evolution of measurement devices and experimental protocols).
- A second group is related to the means used to collect data. Information related to these criteria is typically included in the Materials and Methods section in life science papers, which thoroughly describes the experimental protocol and materials used. Some methods may be known to be less accurate than others, but still be chosen for practical considerations.
- A third group is related to statistical procedures: presence of repetitions, uncertainty quantification, elaboration of an experimental design.

This list of criteria can be reduced or enriched, according to the relevant features to evaluate reliability and the available information in the data sources. Examples of criteria may be found in Buche et al. (2021b) for the milk microfiltration process. An excerpt of this list can be found in Figure 2.21, where a reliability criterion assessment is linked to each modality of a given criterion by domain experts. For example, in Figure 2.21, the modality “3 and more repetitions” that is linked to *MF-Number of repetitions - Compositional analysis* criterion is assessed as “very reliable.” The assessments provided by a domain expert are used in Figure 2.22 to assess the reliability of Jørgensen et al. (2016). It is the role of the annotator of the external source to associate values with criteria during the annotation process. Qualitative assessments associated with modalities of reliability criteria are merged to provide a unique global assessment represented as a couple (low expectation, high expectation). Expectations vary in the interval [1,5], where 1 represents a low reliability and 5 a high reliability. In Figure 2.22, the merging of the nine assessments corresponds to the couple [4.7, 4.99], meaning that the reliability is high and there is no contradiction between criteria as low and high expectations are close. Fuzzy set and belief functions theories are used (i) to represent the individual assessments and (ii) to merge them to compute the global one. For more details about this modeling, see Destercke et al. (2013).

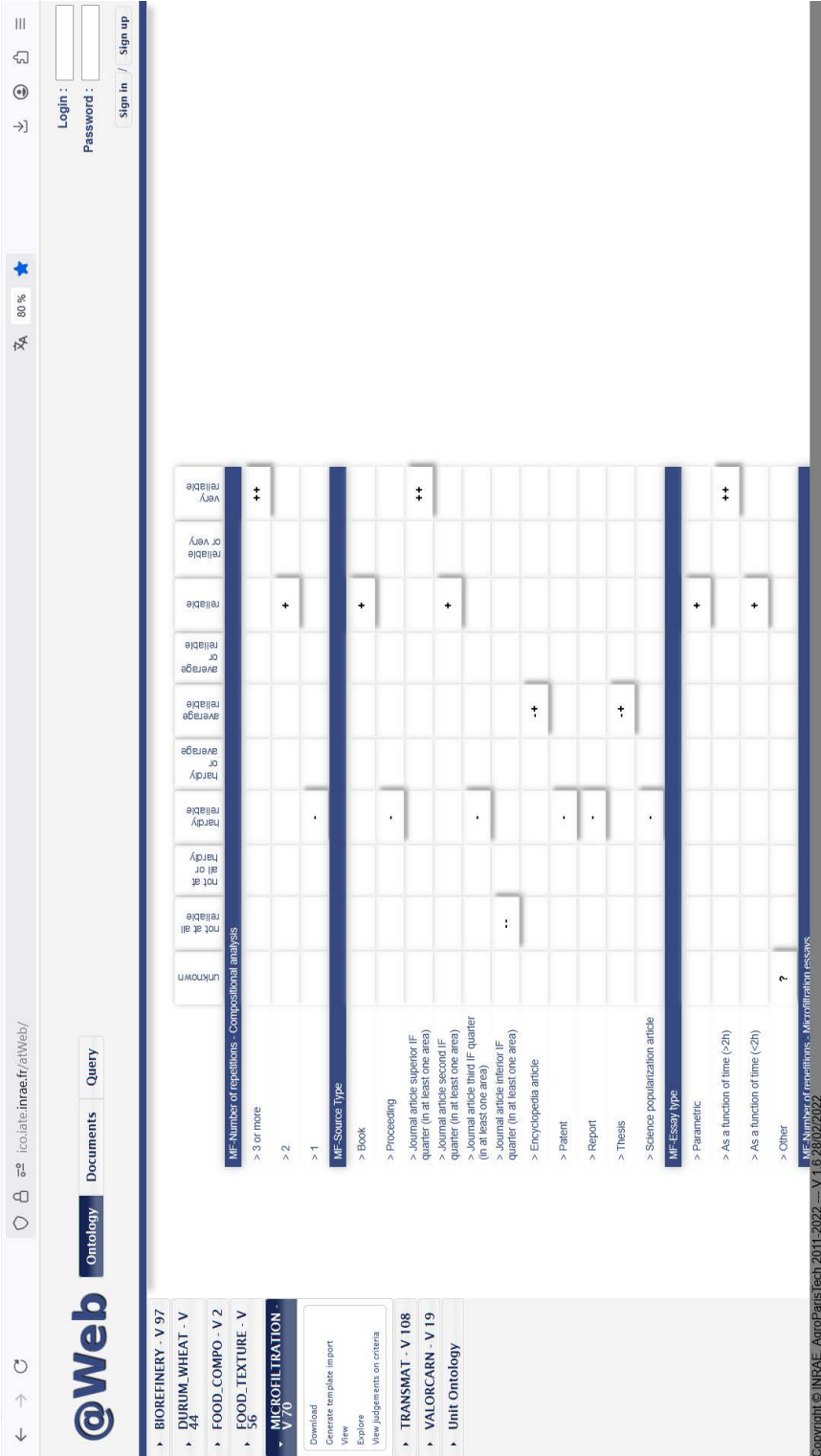


Figure 2.21. Screenshot of @Web showing examples of reliability criteria assessment associated with an external data source in the field of microfiltration made by domain experts.



Figure 2.22. Screenshot of @Web showing a reliability assessment associated with the external data source Jørgensen et al. (2016). MF: microfiltration, IF: impact factor, UTP: uniform transmembrane pressure, GP: graded permeability.

►► 2.3 Outcomes

Reuse of data stored in the knowledge graphs presented in this chapter will be found in Chapter 4. In this section, we will present several resources (especially the location of ontologies and available data papers) and statistical indicators associated with the knowledge graphs.

2.3.1 PO² resources and statistical indicators

The PO² ontology can be found on Recherche Data Gouv.²³ TransformON is also available on both Recherche Data Gouv²⁴ and AgroPortal.²⁵

The current version of the PO² knowledge graph includes:

- 108 projects
- 1,634 itineraries
- 6,850 steps
- 18,674 tables of observations
- 221,635 observations
- 22,988,141 RDF triples

To date, three data papers have been published to describe data stored in the PO² knowledge graph. The type of data is provided in the data paper title column in the following table (Table 2.1).

Table 2.1 Data papers published to describe data stored in the PO² knowledge graph

Data paper title	Data paper DOI
Biocomposites from poly(3-hydroxybutyrate-co-3-hydroxyvalerate) and lignocellulosic fillers: Processes stored in data warehouse structured by an ontology	https://dx.doi.org/10.1016/j.dib.2022.108191
Datasets on the production routes and the properties of plant powders for manufacturing of green products	https://dx.doi.org/10.1016/j.dib.2024.110787
Dataset for common wheat (<i>Triticum aestivum</i> L.) grain and flour characterization using classical and advanced analyses	https://dx.doi.org/10.1016/j.dib.2025.111375

2.3.2 CAPEX indicators

CAPEX ontology is available on Recherche Data Gouv.²⁶ Currently, two CAPEX knowledge graphs have been created: the first as part of the DOCAMEX association and the second for the INRAE IATE PLANET platform.

Indicators on DOCAMEX knowledge graphs are not provided as they are private data. However, to date, 20 protected designation of origin (PDO) cheese sectors

23. <https://doi.org/10.15454/XSVVBW>

24. <https://doi.org/10.57745/DWX7W6>

25. <https://agroportal.lirmm.fr/ontologies/TRANSFORMON>

26. <https://doi.org/10.57745/SEJP1B>

representing more than 70% of French production use the tool and manage several dozen decision trees per sector. The current PLANET knowledge graph includes more than 57 decision trees.

The CAPEX OWL ontology, an example of a decision tree, and associated efficiency indicators are available in Buche et al. (2022). Indicators about the use of decision trees are available in Chapter 4 (Section 4.3: Outcomes).

2.3.3 @WEB indicators

Three ontologies have been created using naRyQ task ontology and are available on Recherche Data Gouv: BIOREFINERY,²⁷ TRANSMAT,²⁸ and MICROFILTRATION.²⁹ They are also distributed on AgroPortal. Links on resources distributed on AgroPortal are available in the dataset metadata published on Recherche Data Gouv for each ontology. In the following table (Table 2.2), the number of n-ary relations indicates the richness of the ontological model as the number of instances informs the quantity of available data in each knowledge graph. The BIOREFINERY and MICROFILTRATION knowledge graphs contain itineraries of biomass transformation and input biomass and final product characterization. The TRANSMAT knowledge graph includes only food and packaging characterizations.

Table 2.2 Domain ontologies created using naRyQ task ontology

Domain ontology	Number of n-ary relations	Number of instances of n-ary relations
BIOREFINERY	92	149,213
TRANSMAT	64	39,751
MICROFILTRATION	20	144,239

Seven data papers have been published to describe data stored in the @WEB knowledge graph. The type of data is provided in the data paper title column in the following table (Table 2.3).

Table 2.3 Data papers published to describe data stored in the @WEB knowledge graph

Data paper title	Data paper DOI
CO ₂ solubility and composition data of food products stored in data warehouse structured by an ontology	https://dx.doi.org/10.1016/j.dib.2023.108950
Partial n-ary relation instances on food packaging composition and permeability extracted from scientific publication tables	https://dx.doi.org/10.1016/j.dib.2022.108000
Milk microfiltration process dataset annotated from a collection of scientific papers	https://dx.doi.org/10.1016/j.dib.2021.107063
Food packaging permeability and composition dataset dedicated to text-mining	https://dx.doi.org/10.1016/j.dib.2021.107135

27. <https://doi.org/10.15454/X2MOWO>
28. <https://doi.org/10.15454/NK24ID>
29. <https://doi.org/10.15454/5MQMKG>

Data paper title	Data paper DOI
Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders	https://dx.doi.org/10.1016/j.dib.2020.106430
Consumer perception data and scientific arguments about food packaging functionalities for fresh strawberries	https://dx.doi.org/10.1016/j.dib.2018.09.034
CO ₂ and O ₂ solubility and diffusivity data in food products stored in data warehouse structured by ontology	https://dx.doi.org/10.1016/j.dib.2016.04.044

2.3.4 Reliability model for the empirical validation of external data sources

The conceptual reliability model presented in this chapter has been empirically validated twice in decision support systems dedicated to food and bioresource processing: (1) to assess reliability of aggregated external sources of information used to rank pretreatment itineraries to produce glucose from lignocellulosic biomass according to a green indicator (Lousteau-Cazalet et al., 2016), and (2) to predict performances of milk micro-filtration under a wide range of operating conditions and membrane technologies (Baudrit et al., 2022). Interested readers may refer to those publications for detailed explanation. Figure 2.23 provides an illustration from Lousteau-Cazalet et al., (2016)

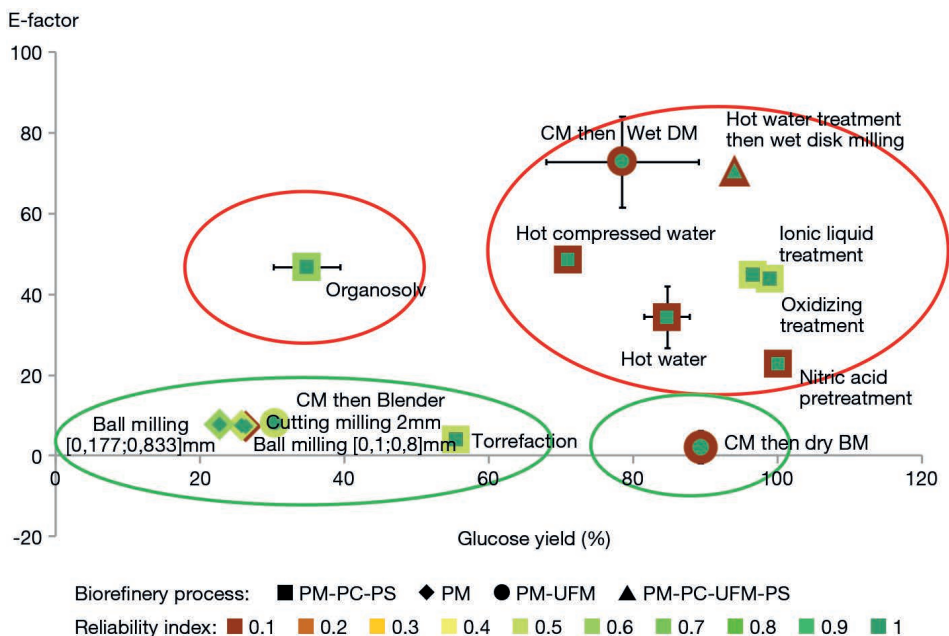


Figure 2.23. Reliability assessment associated with Efactor computation with rice straw pretreatments for best experiment of scientific articles. BM: ball milling, CM: cutting milling, DM: disk milling, PC: physical/chemical treatment, PM: pre-milling, PS: press and separation, UFM: ultrafine milling.

that shows two categories of environmental impact indicators associated with ligno-cellulosic pretreatments. The first one (green circles and squares) represent experimental data from papers associated with a good reliability index assessed without contradiction between metadata, and the second one (red/green circles and squares) represent experimental data from papers associated with a good reliability index assessed with contradiction between metadata. In Figure 2.23, reliability indexes vary in the interval $[0.1, 1.0]$, where 0.1 represents low reliability and 1.0 high reliability.

► 2.4 Conclusion and resources

This chapter tackled four research questions:

- Q1: How can we represent in terms of generic knowledge and associated experimental data (i) a process comprising the steps through which raw materials (input components) are transformed into final products (output components) and (ii) observations associated with steps and components?
- Q2: How can we represent expert and operational collective know-how that enables the formulation of recommendations on technological actions that may help maintain product (resp. observation) quality or correct a product (resp. observation) quality defect at the scale of a given food or biobased product processing operation?
- Q3: How can we represent data tables associated with external sources using the same vocabulary as that used for indexing the internal data?
- Q4: How can we assess the reliability of external data sources and take this assessment into account in decision support systems?

Four solutions were put forward to answer these questions and compared to the current state of the art. For each solution, the software workflow implementation was presented and illustrated. Available open resources (data, software, video, courses) were listed. Main outcomes obtained using those solutions were been presented. All have very high technology readiness levels and they are all being used in production. With regard to task ontological-based solutions – namely PO^2 , CAPEX and @Web – we focused our efforts on ensuring that our three ontologies met the real-world use cases proposed by users. Now that we have reached a stable version with our models, we believe it would be highly beneficial to implement ontological design patterns in this work to facilitate their reuse.

A complete method to semi-automatically extract data from tables of external sources based on the naRyQ task ontology is presented in Chapter 3. Multicriteria decision support using the content of those knowledge graphs is presented in Chapter 4.

Ongoing research focuses on the interoperability between two of those solutions (PO^2 and CAPEX) as well as MAKEBOOK (Baudrit et al., 2025), another solution that can be used to represent generic knowledge using concepts and relations thanks to mind-mapping schemas. MAKEBOOK can be instantiated for a given application domain, which in our case is biomass processing and characterization. It is currently being extended to use PO^2 /TransformON concepts and relations to facilitate interoperability with PO^2 and CAPEX knowledge graphs (see Figure 2.24).

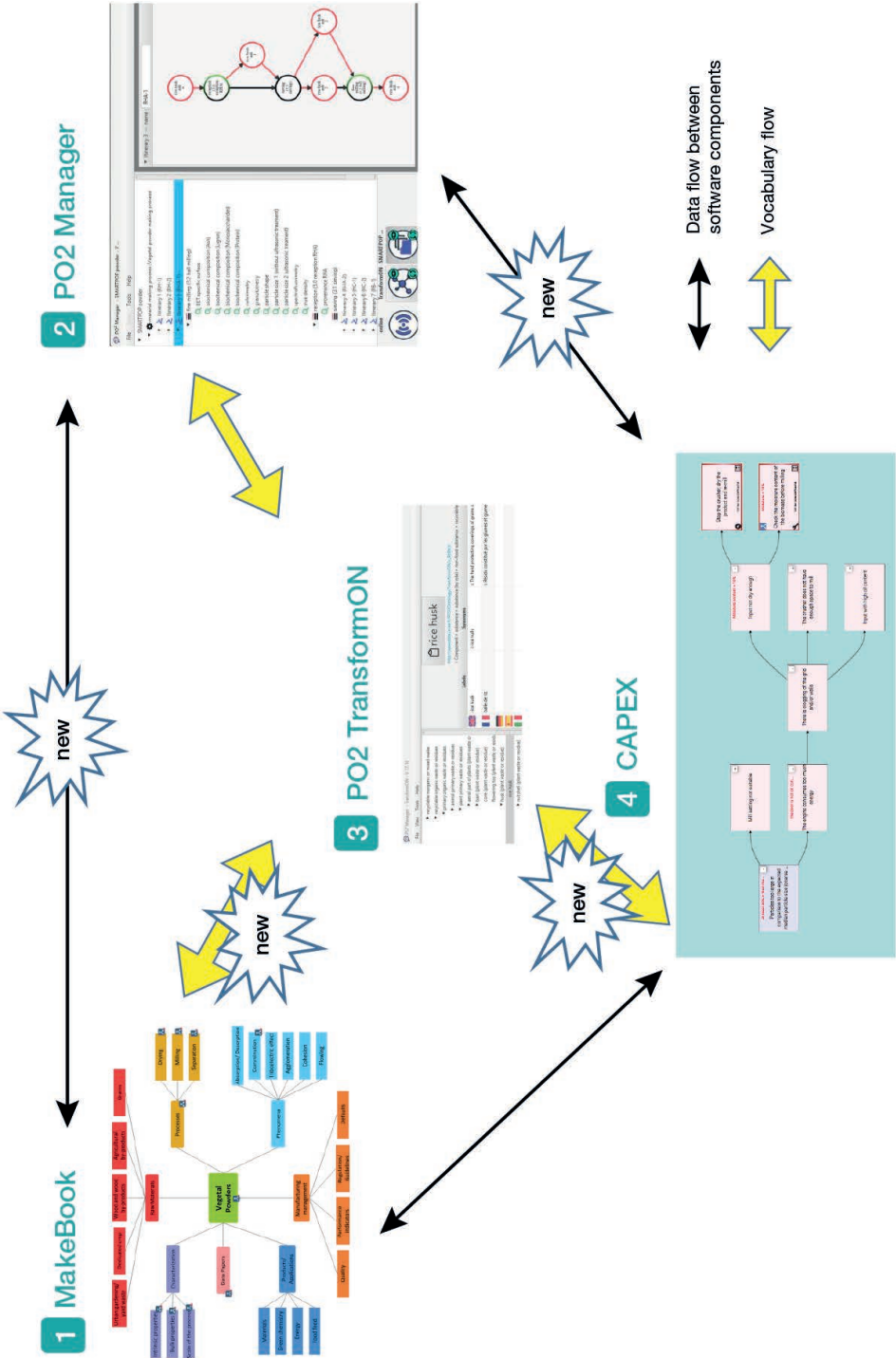


Figure 2.24. Interoperability between software components.

Available resources

Available resources				
Research question	Proposed solution	Available resources		
		Main scientific papers	Ontology, data papers and datasets	Software codes
Q1	PO ²	Ontological model (https://hal.science/hal-03662183v1 , https://hal.inrae.fr/hal-04197618)	See Section 2.3: Outcomes	SPO ² Q (https://hal.science/hal-04501660), PO ² Engine (https://hal.science/hal-04387669), PO ² Manager (https://hal.science/hal-04313202)
			Smartpop project consultation with PO ² Manager https://hal.inrae.fr/hal-05678607v1 Querying Smartpop PO ² project https://hal.inrae.fr/hal-05678619v1	7 courses to learn how to use PO ² Manager and SPO ² Q (in French) Introduction https://hal.inrae.fr/hal-05315840v3 5 courses https://hal.inrae.fr/hal-05315873v3 https://hal.inrae.fr/hal-05315863v3 https://hal.inrae.fr/hal-05315887v3 https://hal.inrae.fr/hal-05315902v3 https://hal.inrae.fr/hal-05315911v3 Guide https://hal.inrae.fr/hal-05315925v3
Q2	CAPEX	Ontological model (https://hal.science/hal-04086712), knowledge elicitation method (https://hal-lirmm.ccsd.cnrs.fr/lirmm-02162114)		CAPEX https://hal.science/hal-04957075v1

Research question	Proposed solution	Available resources				Courses and technical reports
		Main scientific papers	Ontology, data papers and datasets	Software codes	Videos	
Q3	@Web	Ontological model (https://hal-lirmm.ccsd.cnrs.fr/lirmm-00642899)			Updating symbolic concepts in OTR https://hal.science/hal-05191864v1	User manual (in French) https://hal.science/hal-05196134v1 CSV files for ontology import (in French) https://hal.science/hal-05196137v1
Q4	Reliability assessment in @Web	Reliability assessment model (https://hal.science/hal-01267940v1)	Examples of reliability criteria defined for milk microfiltration (https://hal.inrae.fr/hal-03227408v1)	R Code available at https://cran.r-project.org/src/contrib/Archive/belief	Reliability criteria quotation: https://hal.science/hal-05191871v1	Technical report on reliability assessment implementation in @Web platform https://hal.inrae.fr/hal-05682902v1 Technical report on querying with reliability assessment implementation in @Web platform https://hal.inrae.fr/hal-05682908v1

Chapter 3

Methods for Ontology-Guided Extraction of Experimental Data in Scientific Articles

Martin Lentschat, Patrice Buche, Julien Cufi, Luc Menut, Mathieu Roche

► 3.1 Introduction

In this chapter, we will discuss methods to extract experimental data from scientific articles. The content of research publications (e.g., results, claims, technical parameters, tools, and methodologies) is currently underexploited. A system that could semi-automatically extract relevant information and create a structured representation of knowledge, following standard formats used in knowledge bases, would be very helpful. Such a system would, for instance, make it possible to enrich scientific literature meta-analyses or foster study replication, and could later be used by expert and decision support systems.

The main challenge in achieving this aim is the heterogeneity of data present in the literature. First, scientific articles are written in natural language, which differs significantly from a computer-structured representation. The experimental data of interest can also be presented in various contexts and forms (e.g., numerical values, entities of interest). For instance, the methodology and materials used are described in text in the first part of the document, while the related experimental results are presented in table format, with additional details provided in the discussion section.

Representing knowledge using n-ary relations makes it possible to link pieces of information associated with experimental data and the metadata that describe that information's context, given that these details are scattered throughout the scientific article in text format. In this chapter, we will use the naRyQ task ontology presented in Chapter 2 to represent n-ary relations. The naRyQ task ontology will be specialized for a given domain and called the ontological and terminological resource (OTR) in this chapter. The excerpt of naRyQ in the matter transfer domain, called TRANSMAT,³⁰ was presented in Chapter 2 (Section 2.2.3).

The extraction of n-ary relations representing gas permeability properties of food packaging will be used as an example. Example 3.1 (below) shows how its argument instances can appear in a document.

30. MATter TRANSfer ontology, available at <http://doi.org/10.15454/1.430814846357855E9>

Example 3.1. Appearance of argument instances in a text:

“The permeability of ethylene vinyl alcohol (EVOH) films was measured with the ASTM D95-96 method at $25\pm1^\circ\text{C}$. The film had a thickness of $15\mu\text{m}$ and showed optimal barrier properties with a permeability to oxygen of $4.34\pm1.03\text{ cm}^3\mu\text{m}^{-2}\text{ d}^{-1}\text{ kPa}$. This measurement was obtained at a constant RH of 85.0%.”

Figure 3.1 illustrates an instance of an n-ary relation representing the information presented in Example 3.1. The n-ary relation instance is structured around a relation concept (see center of Figure 3.1) linked to its component arguments (e.g., permeability value, packaging involved, experimental conditions).

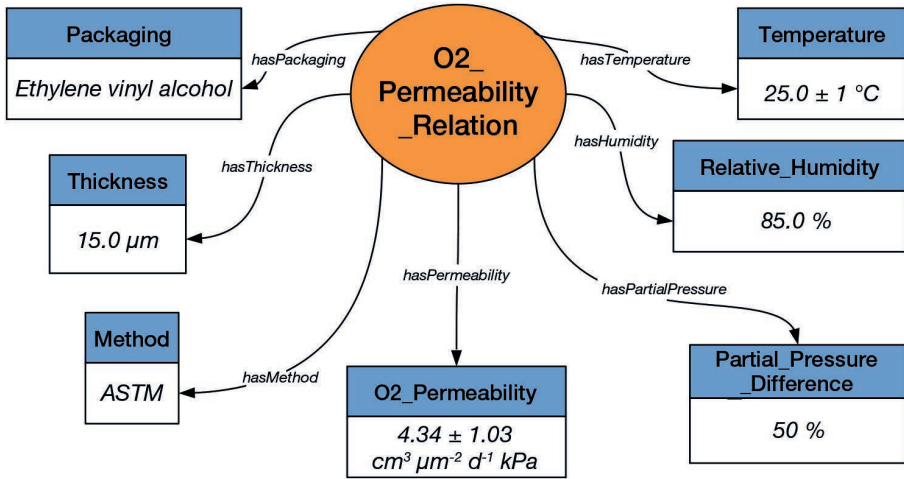


Figure 3.1. Instance of an oxygen permeability relation.

However, Example 3.1 is a fabricated example. Typically, most argument instances are scattered across the document and need to be linked together to form relation instances (see Section 3.2.3 for more details). Argument instances that appear as word(s) are called symbolic arguments, and pairs of numerical values and units of measurement form quantitative arguments. The main obstacles to identifying argument instances and reconstructing relations are:

- lexical variations (e.g., acronyms, new units of measurement or packaging names), which means the OTR vocabulary must be extended to cover most cases (see Section 3.2.1);
- presence of false positives (e.g., not all temperatures in a text relate to oxygen permeabilities) that need to be filtered or discarded (see Section 3.2.2);
- dispersion of argument instances (see Section 3.2.3);
- lack of information, as authors may omit some details or consider them implicit (e.g., in the annotated corpus of Lentschat et al. (2021a), most of the argument instances for partial pressure differences are not present). This underlines the need for methods that would enable partial relations to exist (see Section 3.2.3).

The main research question addressed in this chapter is:

- Q1: Is it possible to (semi-)automatically extract targeted data in scientific publications in such a way as to be able to exploit it in downstream systems?

Focusing on the extraction of experimental knowledge from scientific articles, in the form of n-ary relation instances, falls under the area of information and relations extraction, and more specifically, extraction of complex relations in specialized domains. This specific question was previously addressed using data mining approaches (Berrahou et al., 2017), guided by the naRyQ ontological and terminological resource (OTR) (Touhami, 2011) presented in Chapter 2 (see Chapter 2.2.3 for more details). This approach created a two-step hybrid method to automatically find targeted data defined as instances of n-ary relations belonging to the OTR.

The aim of this approach is to identify frequent patterns in the corpus involving arguments of n-ary relations belonging to the OTR. Classification algorithms are employed to identify sentences that may contain arguments. Once applied to a large corpus, the most frequent patterns are extracted to find instances of arguments belonging to n-ary relations appearing in scientific articles. These can constitute partial n-ary relation instances that would then need to be linked together.

However, this approach is limited by the large number of potential patterns obtained. Frequency alone does not necessarily identify the ones that are most efficient or reliable. Also, the case study presented above exemplifies a common problem in text data extraction: the presence of false positives that appear as data of interest (e.g., temperature values in the text are often related to something other than the experimental condition of gas permeability measurements). Relevant argument instances cannot be distinguished from false positive ones through pattern matching. This specific issue will be addressed in the approach presented here.

The main hypothesis of the approach presented in this chapter consists in breaking down the process of extracting instances of n-ary relations into two phases. Phase I focuses on identifying and extracting entity argument instances found in scientific articles relating to the n-ary relations of interest. Phase II seeks to link the previously extracted entities, regarded as relation argument instances, to reconstitute instances of n-ary relations. This involves creating a workflow that includes information extraction and relation extraction tasks.

Previous research has demonstrated that different sections of scientific articles contain significantly different information (Cohen et al., 2010; Mao et al., 2003; Shah et al., 2003). For instance, Hofmann et al. (2009) determined that the abstract and discussion sections were the most favorable to relevant keyword discovery. Hence, the structure of scientific articles must be taken into account in our representation of argument instances to evaluate the relevance of automatically extracted argument instances.

Scientific articles also make extensive use of acronyms, which is important to recognize as they may be relevant information and entities. Some acronyms are common and shared by all authors in a field (e.g., NLP = natural language processing), some may be more specialized and related to a particular subject (e.g., OTR = ontological and terminological resource), and some may only be defined for the particular needs of a study (e.g., SciPuRe = scientific publication representation). Because of these widely different situations, the use of an existing resource (e.g., Acronym Finder³¹) cannot cover all the cases found in a publication. The main way of approaching this problem is to recognize potential acronyms through a set of heuristics, such

31. <https://www.acronymfinder.com/>

as identifying terms composed of capital letters and not belonging to a standard dictionary (Ehrmann et al., 2013; Wren and Garner, 2002). Once these potential acronyms have been identified, the aim is to find their extended forms in texts, often present once alongside their acronyms. This approach can be further refined using linguistic rules and syntactic analysis (Pustejovsky et al., 2001), statistical measures (Jain et al., 2007; Okazaki and Ananiadou, 2006), or machine learning (Xu and Huang, 2005). In our work, we designed a simple and agnostic approach using regular expressions and syntactic rules (Lentschat et al., 2020) to have a method able to use any domain OTR and only extract acronyms corresponding to its concept labels.

Experimental fields also make extensive use of units of measurement; recognizing these units is essential for information extraction (Jessop et al., 2011a, 2011b). Some units are common (e.g., °C, km/h) but others are much more complex, multidimensional, and mostly employed for the needs of the field studied (e.g., $\text{cm}^3 \mu\text{m m}^{-2} \text{d}^{-1} \text{kPa}$). Recognition of these units and any variations (e.g., $\text{cm}^3 \mu\text{m m}^{-2} \text{d}^{-1} \text{kPa} == \text{cm}^3 \cdot \text{m}^{-2} \cdot \text{s}^{-1} \cdot \text{bar} \cdot \mu\text{m}$) is an important issue (Foppiano et al., 2019). Resources listing units of measurement do exist, but little research tackles the extraction of variants of units of measurement. The existing literature (Cohen et al., 2003; Van Assem et al., 2010) mainly uses similarity measures to associate existing units of measurement and variants. In the approach presented in this chapter, we use an existing method (Berrahou et al., 2015) for this task.

Data is also present in tables included in articles. Recent work in the field of data extraction from tables focuses on recognizing the concepts represented and linking them to the data. Most approaches use external resources to guide the categorization of identified data and relations. Among existing systems, TabEL (Bhagavatula et al., 2015) uses concept co-occurrences observed on Wikipedia to assign data to their probable category(ies) in DBpedia, while TableMiner⁺ (Zhang, 2017) uses multiple knowledge bases (e.g., WordNet, Wikipedia, Freebase, DBpedia) to categorize data and exploit contextual information (e.g., the table caption) to improve disambiguation. MTab (Nguyen et al., 2019) tackles the SemTab 2019 challenge of aligning table data with a knowledge base (e.g., DBpedia) by using structural information, first estimating the column category and then resolving ambiguities.

The method we chose (Buche et al., 2013a; Hignette et al., 2009) is based on the use of an OTR dedicated to a given application domain, which is not possible with agnostic state-of-the-art methods. It was selected for its performance, its robustness against the different data formats found in the tables, and its compatibility with the OTR we use. This method also creates an annotation of the tables that can be directly exploited by our method and follows a multistep process similar to that of manual annotation. It produces good results and has been employed (Wolf et al., 2018) as a tool to assist meta-analyses of experimental results by extracting data from tables in scientific papers.

The pipeline detailed in this chapter aims to extract knowledge and represent it as n-ary relations. It is guided by the use of a naRyQ OTR. This enables the recognition of symbolic and quantitative arguments (i.e., information) thanks to the lexical component and representation of knowledge in the form of n-ary relations (i.e., relations between n arguments). These relations can then be instantiated in a knowledge base and used by downstream processes such as expert decision systems.

Our use case concerns food packaging characteristics (i.e., mainly gas permeabilities), although our method was designed to be adaptable to any domain represented in a

naRyQ-OTR. Details on the datasets used (Lentschat et al., 2021a, 2021c), our approaches (Lentschat et al., 2020, 2022a, 2022b), and implementation³² can be found online.

3.2 Proposed solutions

In this section, we present the process of extracting n-ary relation argument instances guided by an OTR. This method is thought to be generic and relies solely on an OTR that describes an experimental domain, defines the types of n-ary relation arguments to be extracted, and guides the extraction (see Figure 3.2). Its main contribution consists in two multicriteria representations: scientific publication representation (SciPuRe) associated with n-ary relation argument instances and scientific table representation (STaRe) associated with n-ary relation instances (see Sections 3.2.1 and 3.2.3), providing rich features actionable in both phases of our pipeline (see Figure 3.2).

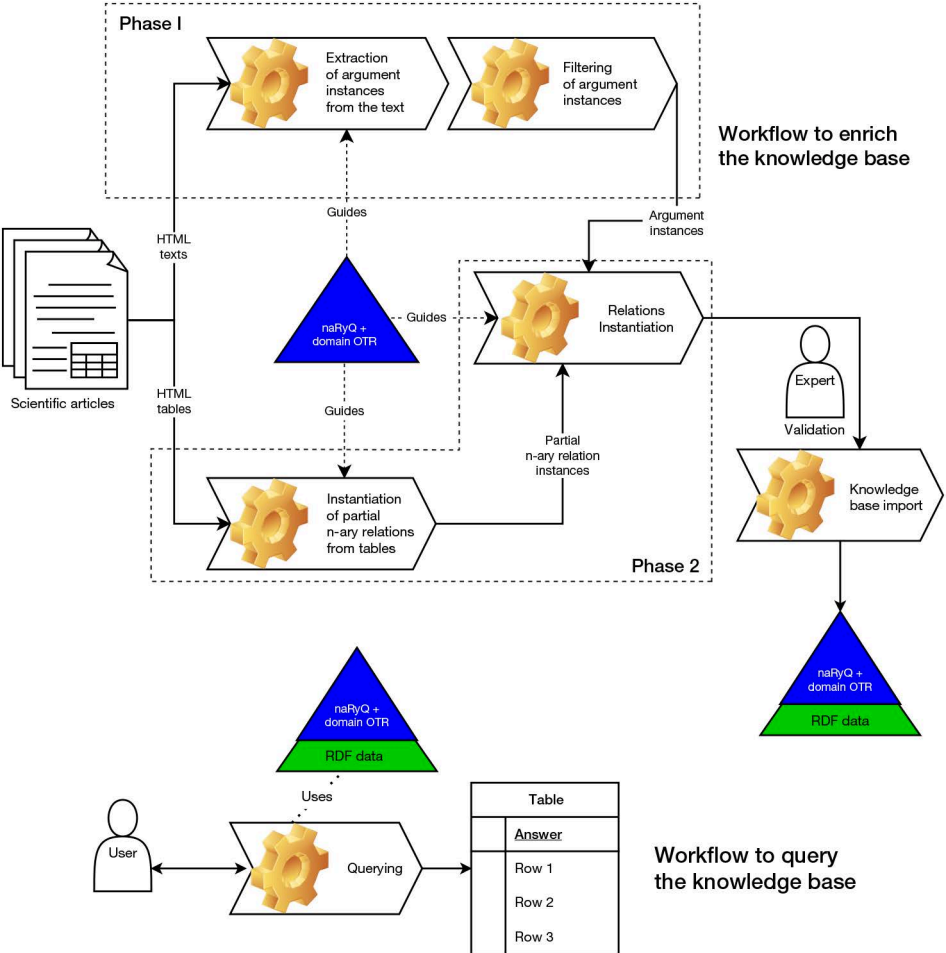


Figure 3.2. General workflow of scientific data extraction and querying.

32. <https://gitlab.softwareheritage.org/eskode/ARTEXT4LOD>

During Phase I, we use a domain naRyQ OTR to guide the extraction of n-ary relation argument instances. The OTR defines n-ary relations in a given domain and their component arguments, while also including a terminological component to detect terms relating to argument instances in documents. It carries out the augmentation of the OTR vocabulary coverage and identifies argument instances along with their representation features, and then measures the relevance of argument instances to filter out false positives (see details in Section 3.2.1).

In Phase II, the aim is to reconstitute instances of n-ary relations (see details in Section 3.2.2). Argument instances making up relation instances are found in the text and tables of each document. The reconstitution of n-ary relations instances is achieved with the partial relations found in document tables (Hignette et al., 2009), completed by a selection of text argument instances using a multicriteria approach. Indeed, whether an argument instance belongs to a relation instance is not explicit in the texts. Additionally, while some argument instances only belong to one relation instance (e.g., the permeability value), other argument instances may belong to different relation instances (e.g., the temperature of the chamber is used in an oxygen, carbon dioxide, and dihydrogen monoxide permeability relation).

3.2.1 Extraction and representation of text argument instances

During Phase I, argument instances in the text are extracted. First, coverage of the OTR vocabulary is increased and then a representation, called scientific publication representation (SciPuRe) is built by aggregating ontological, structural, and lexical features. This representation is used to filter relevant argument instances, and later during Phase II to reconstruct relations. The Phase I process is detailed below and illustrated in Figure 3.3.

SciPuRe: scientific publication representation

During Phase I, we build a representation, called scientific publication representation (SciPuRe) (Lentschat et al., 2020), of extracted argument instances. It describes their characteristics and context of appearance. SciPuRe integrates a set of ontological, lexical, and structural features, following important criteria for representing information (Boyce et al., 2007): discriminating differences, identifying similarities, describing elements accurately, and minimizing ambiguity. Representations of extracted data in information extraction sometimes focus on the specific needs of the study (e.g., representation of quantitative data as in Hao et al., 2017) and exploit mainly lexical features (Wood et al., 2021; Zaenen et al., 2010) or an external resource (e.g., a knowledge base as in Franco-Salvador et al., 2014). Our representation offers a high degree of descriptive richness for argument instances. As such, the argument instances can be described according to several criteria, while remaining domain agnostic and adapted to different data formats, both symbolic and quantitative. This approach is situated in the smart data paradigm (Duong et al., 2017; Zeng, 2017).

SciPuRe contains three categories of features, or information common to all argument instances:

- **Ontological features:**

- *Target* indicates the generic type of the argument (i.e., the highest concept in the ontology hierarchy that is part of an n-ary relation signature, such as “packaging” or “temperature”).

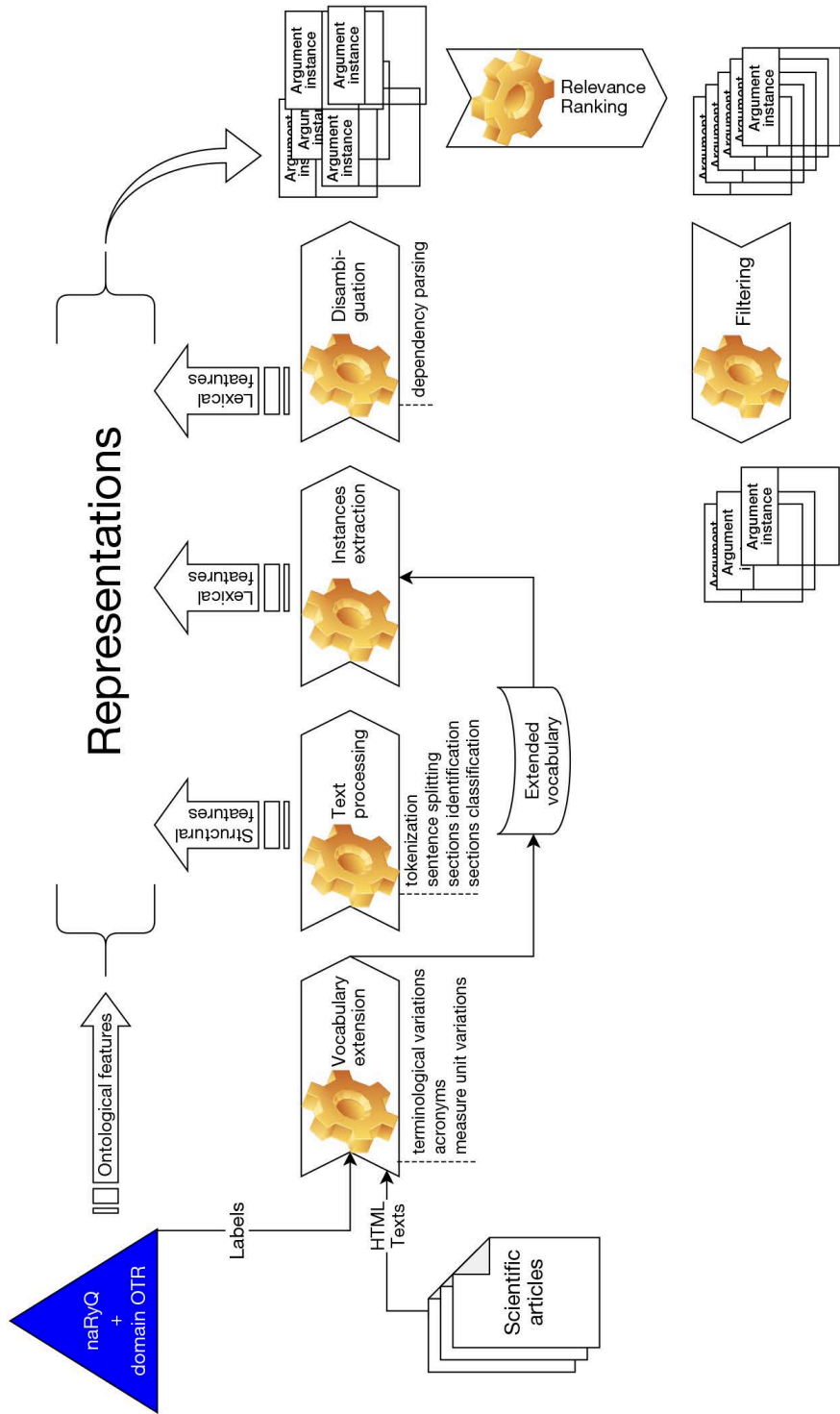


Figure 3.3. Phase I: textual argument instance extraction and representation.

- *Node* specifies the subconcept that the instance represents (i.e., the specific OTR concept associated with the label used for instance recognition or disambiguation).
- **Lexical features:**
 - *Original_Value* contains the text corresponding to the extracted argument instance.
 - *Attached_Value* corresponds to the term(s) associated with the instance and used in the disambiguation step.
- **Structural features:**
 - *Sentence* contains the text of the argument instance sentence.
 - *Window* provides a larger context with the previous and next sentences.
 - *Segment* takes into account the structure of a scientific article (i.e., sections and subsections).
 - *Document* provides the article’s references (e.g., title or DOI).

SciPuRe is illustrated in Examples 3.3a and 3.3b.

Example 3.3a. SciPuRe representation of the packaging argument instance in the n-ary relation of Figure 3.1, from Lentschat et al. (2020).

Ontological	Target	Packaging
	Node	Ethylene_Vinyl_Alcohol
Lexical	Original Value	<i>“ethylene vinyl alcohol films”</i>
	Attached Value	<i>“ethylene vinyl alcohol films”</i>
Structural	Sentence	<i>“The permeability of ethylene vinyl alcohol films (EVOH) ... at 25±1 °C.”</i>
	Window	[..., <i>“The permeability of ... at 25±1 °C.”</i> , <i>“The film had a thickness ... kPa.”</i>]
	Segment	Introduction
	Document	10.xxxx/xxxxxx-xxx-xxxx-x

Example 3.3b. SciPuRe representation of the oxygen permeability argument instance in the n-ary relation of Figure 3.1, from Lentschat et al. (2020).

Ontological	Target	Oxygen_Permability
	Node	Oxygen_Permability
Lexical	Original Value	<i>“4.34±1.03 cm³ µm⁻² d⁻¹ kPa”</i>
	Attached Value	<i>“permeability to oxygen”</i>
Structural	Sentence	<i>“The film had a ... permeability to oxygen of 4.34±1.03 cm³ µm⁻² d⁻¹ kPa.”</i>
	Window	[<i>“The permeability of ... at 25±1 °C.”</i> , <i>“The film had a ... permeability to oxygen of 4.34±1.03 cm³ µm⁻² d⁻¹ kPa.”</i> , <i>“This measurement ... RH of 85.0%.”</i>]
	Segment	Introduction
	Document	10.xxxx/xxxxxx-xxx-xxxx-x

Extraction of argument instances

Phase I uses an OTR to extract the argument instances of the n-ary relations of interest in the text of the documents. Scientific texts are first preprocessed using pretrained language models and their associated libraries. This common operation brings out text features such as sentence boundary detection, tokenization (i.e., identifying word or subwords), part-of-speech tagging (e.g., verb, noun, determiners), or syntactical analysis. Some of the best-known language models used for these operations are BERT (Devlin et al., 2018) from Google, Stanza (Qi et al., 2020) from Stanford University, or the models distributed by the company spaCy.³³ These models are typically trained on large corpora of generic or specialized texts (e.g., OntoNotes Release 5.0³⁴ or CoNLL-2003 [Tjong Kim Sang and De Meulder, 2003]).

We have broken down the extraction of argument instances (shown in Figure 3.3) into four steps:

- The Variation Extraction step first increases the coverage of the OTR used by searching for terminological variations of OTR labels, acronyms, and new units of measurement. This identifies as many terms as possible to be used as instances of symbolic arguments or to disambiguate instances of quantitative arguments. The tool FASTR (Bourigault and Jacquemin, 1999) is used to extract terminological variations. It relies on linguistic rules and can handle multiword terms. Acronyms are recognized using regular expressions and aligned with OTR labels based on the Dice score (Dice, 1945). New units of measurement are added to the OTR using an existing method (Berrahou et al., 2015). In Example 3.1, the retrieved terminological variations and new units of measurement are:

- “*permeability to oxygen*” is a variation of the OTR label “*oxygen permeability*”
- “*RH*” is an acronym of the OTR label “*relative humidity*”
- “*cm3 μm m-2 d-1 kPa*” is a new unit of measurement linked to the existing unit of measurement in the OTR “*cm3.m-2.s-1.bar.μm*”

- The Text Processing step uses the SciSpacy language model. It covers:

- tokenization that identifies the lowest-level units at which the program is functioning
- sentence splitting, information used to inform SciPuRe features
- section identification, based on the document structure (i.e., using the HTML tags) and grouped into categories (e.g., “Conclusion and Discussion” is in the same category as “Conclusion and Future Work”) in SciPuRe

- The Entity Extraction and Disambiguation steps build on the extended OTR terminological component to identify document terms related to instances of symbolic and quantitative arguments.

Example of entity identification in Example 3.1:

The permeability of **ethylene vinyl alcohol (EVOH)** films was measured with the **ASTM D95-96** method at **25±1°C**. The film had a **thickness** of **15 μm** and showed optimal barrier properties with a **permeability to oxygen** of **4.34±1.03 cm3 μm-2 d-1 kPa**. This measurement was obtained at a constant **RH** of **85.0%**.

Key: **symbolic concept label** **quantity concept label** **unit of measurement** **numerical value**

33. <https://spacy.io/>

34. <https://catalog.ldc.upenn.edu/LDC2013T19>

Associations between numerical values and units of measurement (e.g., “ μm ” is the unit of measurement for “15” in the text above) are retrieved by analyzing the syntactic dependencies using the Stanza language model ParTUT,³⁵ since it performed better than SciSpacy on this specific task. The same process is used to associate each ambiguous unit of measurement (e.g., % is ambiguous since, as opposed to °C, which always represents a temperature, % does not carry its quantity concept) with a term describing a quantity and to determine its respective quantitative concept (e.g., if “RH” is syntactically associated with “85.0%” through the preposition “of,” then the argument is Relative_Humidity).

The steps constituting Phase I rely on direct recognition techniques guided by the structure and terminological component of an OTR and pretrained language models. This process has been designed to be generic so it can be used in different domains for the extraction of experimental data, guided by an OTR from that domain.

3.2.2 Filtering relevant argument instances

Following the extraction of argument instances and the construction of their SciPuRe representations, we propose indicators to measure the relevance of the extracted instances. This step is crucial, as not all previously identified argument instances are relevant for the n-ary relations. For instance, in our application domain, only the packaging types under experimentation in the article are relevant, as they form part of instances of n-ary relations of interest, whereas the packaging types cited for comparison are not. Similarly, there is a large number of temperature values in a document, and the ones used as control parameters in the gas permeability measurement should not be confused with those involved in packaging preparation or food sample conservation. The relevance of information is usually defined by how well it satisfies a user’s information need (Cooper, 1971).

We decided to address this question by using SciPuRe features to compute relevance scores for each of the instances extracted in Phase I. The aim was to assign higher relevance scores to instances that will be involved in the final n-ary relations. The hypothesis was that ordering argument instances according to our scores should allow us to filter relevant instances from irrelevant ones. We experimented with lexical, semantic, and combined relevance scores.

Lexical relevance score

To differentiate between relevant argument instances, lexical relevance scores based on term occurrences are proposed. A classical score used in information systems is the term frequency – inverse document frequency (TF-IDF) score. TF is based on the assumption that the most frequent terms in a document are the most important, while IDF assumes that terms that appear only in certain documents are more specific (Salton and McGill, 1986). Inverse categorical frequency (ICF) is a generalization of IDF to any text category (Wang and Zhang, 2010), not only documents. Our approach uses ICF on sections, based on the assumption that some sections will contain relevant argument instances and others will be more prone to false positives (e.g., control parameters are more often present in the Materials and Methods sections and false positives are common in the Discussion or Future Work sections). SciPuRe features

35. https://github.com/UniversalDependencies/UD_English-ParTUT

enable the use of different textual forms (e.g., Original or Attached Value) and contexts (e.g., Document, Segment). This facilitates adaptation to argument specificities, like differences between symbolic and quantitative arguments. Results are discussed in Section 3.3.3 and detailed in a previous publication (Lentschat et al., 2022b).

In Equation 3.1, TF favors the most frequent argument instances in a document. It measures the relative frequency of a term t in a document d :

$$TF(t, d) = \frac{freq(t, d)}{\sum_{t' \in d} freq(t', d)}$$

Equation 3.1. Term frequency.

In Equation 3.2, IDF and ICF represent how specific a certain term t is to a category of text compared to the whole corpus $|D|$. In ICF, document d is changed to another structure (i.e., all Introduction sections are considered one document, Methods are another document, and so on).

$$IDF(t, d) = \log \frac{|D|}{|\{t \in d\}|}$$

Equation 3.2. Inverse document frequency.

SciPuRe offers different levels of representation attached to each argument instance. The term t can be the specific manifestation of the argument (i.e., the Attached Value feature), or all instances of arguments sharing the same specific concept in the OTR (i.e., the Target feature), which allows for considering terminological variations of the same information. The document d (i.e., Document feature) can typically be replaced by the sections (i.e., the Segment feature) under the assumption that different parts of a publication do not contain the same information.

Considering the differences between symbolic and quantitative arguments, we experimented with different scores to measure their relevance. Details are found in Section 3.3.3.

Semantic relevance score

Since the n-ary relation argument instances extracted from scientific documents are intended to be used by experts and advanced decision support systems, their relevance score must also reflect their informative power. This informative power can be considered through the specificity (Harispe, 2014) of the concept.

For instance, “*PE films coated with chitosan*” and “*multilayer film*” are two ways of referring to packaging. However, the second is obviously less specific than the first. In the OTR, Polyethylene_Film is a subconcept of Multilayer_Film, which itself is a subconcept of Packaging. The notion of semantic relevance is therefore based on the subsumption relation of the concepts in the OTR structure.

SciPuRe informs the specific OTR concepts associated with the instance (i.e., the Node feature) and its top concept (i.e., the Target feature). The distance, taken as the number of edges in the OTR graph between the concepts corresponding to Node and Target, is calculated using the OTR concept hierarchy. This hierarchy expresses the instance specificity (Hararay et al., 1965) that we model in Equation 3.3 as the conceptual distance (CD) score.

The CD score measures the relative distance between the Node n and the Target a of the argument instance compared to the maximum distance existing in all subsumption relations starting from the Target a (i.e., an OTR subtree with a as its root). This relevance score is expected to be especially discriminative for instances of symbolic arguments, as they are described with levels of specialization in the OTR (i.e., subsumption relations in quantitative concepts are very rare).

$$CD(n, a) = \frac{1 + \text{dist}(n, a)}{1 + \max(\text{dist}(n', a) : n' \subseteq a)}$$

Equation 3.3. Conceptual distance.

Combined scores

The relevance scores presented above can be used alone or combined. Linear and sequential combinations of scores are used to combine the properties associated with each type of relevance score (note that they must first be normalized on a [0,1] scale to avoid side effects). Combining relevance scores thus makes it possible to combine different sorts of SciPuRe features.

Linear combination

The linear combination (Equation 3.4) adds up different scores $s_1, s_2, \dots, s_i$ by assigning a weight $w_1, w_2, \dots, w_i$ to each. The sum of all weights is equal to 1. We experimented with various weight proportions, which allowed us to combine scores while controlling the impact of each on the final ranking.

$$\text{Linear}(s_1, s_2, \dots, s_i) = \{s_1 \cdot w_1 + s_2 \cdot w_2 + \dots + s_i \cdot w_i \mid \sum w_1, w_2, \dots, w_i = 1\}$$

Equation 3.4. Linear combination.

Sequential combination

The aim of sequential combination is to use different relevance scores in succession. After an initial score has been used to rank all argument instances, a threshold is applied to form a subset of the top argument instances. The subset is then re-ranked using a new score. This method can be applied again using a third score and so on. The expectation behind this approach is that certain argument sets would be more sensitive to different scores, and using them in a sequence means that false positives can be progressively filtered out according to different criteria. Naturally, the chosen order in which scores are applied is important and must be informed by expert knowledge, as does the choice of threshold at each step (e.g., arguments with a large proportion of false positives such as Temperature can benefit from a high threshold while lower ones are recommended for rarer arguments such as Method).

For example, the CD score can be used on Packaging instances to first eliminate the least specific instances that carry no informative power (e.g., instances of film categories or the term “packaging”). Then, a lexical score such as TF will select the most frequently extracted instances from the remaining set.

3.2.3 Reconstitution of n-ary relations

The purpose of Phase II is to reconstitute n-ary relations instances. It uses the partial relation instances extracted from the document tables and combines them with argument instances found during Phase I, filtered according to relevance. This phase relies on the features of SciPuRe and STaRe, a representation of partial n-ary relations that factors some argument instance features of SciPuRe. The process, described in Figure 3.4, starts by selecting potential candidates for each missing argument instance in a partial relation. Candidates are then ranked according to methods relying on SciPuRe and STaRe features. An expert can then select the right argument instance fit to complete the relation.

STaRe

Tables in scientific articles contain structured information that can be extracted in the form of partial n-ary relations. This structure can be used to retrieve n-ary relations, or rather partial relations, as not all the information sought is present in the tables. Using a table and the OTR as inputs, it is possible to automate the annotation of table data (Hignette et al., 2009) and the retrieval of partial n-ary relations. The automatic method used in our approach was compared with the results of manual annotation on a corpus of 31 tables (Lentschat et al., 2021c, 2022c) and produced comparable results as with the state of the art on such tables (i.e., overall F1 score of 0.70). The automatic method used extends an existing one (Hignette et al., 2009) with additional preprocessing (e.g., to handle columns split in two).

Our aim is to supplement partial n-ary relations with argument instances from the text. The relation instances extracted from the tables present different characteristics. These are grouped together in a relation representation called scientific table representation (STaRe). It groups together features describing the argument instances that are part of the n-ary relation and includes ontological and lexical features of the argument instances. STaRe also includes ontological and structural features specific to the partial n-ary relation instance. STaRe thus factorizes the structural features of the argument instances, while indicating the context of the partial relation's appearance and its formalization in the ontology.

STaRe describes relations through two types of features:

- Ontological features
 - Relation indicates the relation concept in the naRyQ OTR
 - Result_Argument designates the result argument (i.e., the argument characterizing the relation) of the relation in the OTR
 - Arguments list the rest of the argument constituting the relation
 - STaRe factorizes SciPuRe features in Result_Argument and Arguments through ontological features (i.e., Target and Node) and lexical features (i.e., Original_Value, which takes the value of the cell content, and Attached_Value, which corresponds to the text of the column header)
- Structural features
 - Table: the table title
 - Caption: the table caption
 - Segment: the section in which the table appeared
 - Document: the DOI of the scientific article

Structural features inform the context of the table and place it in the document.

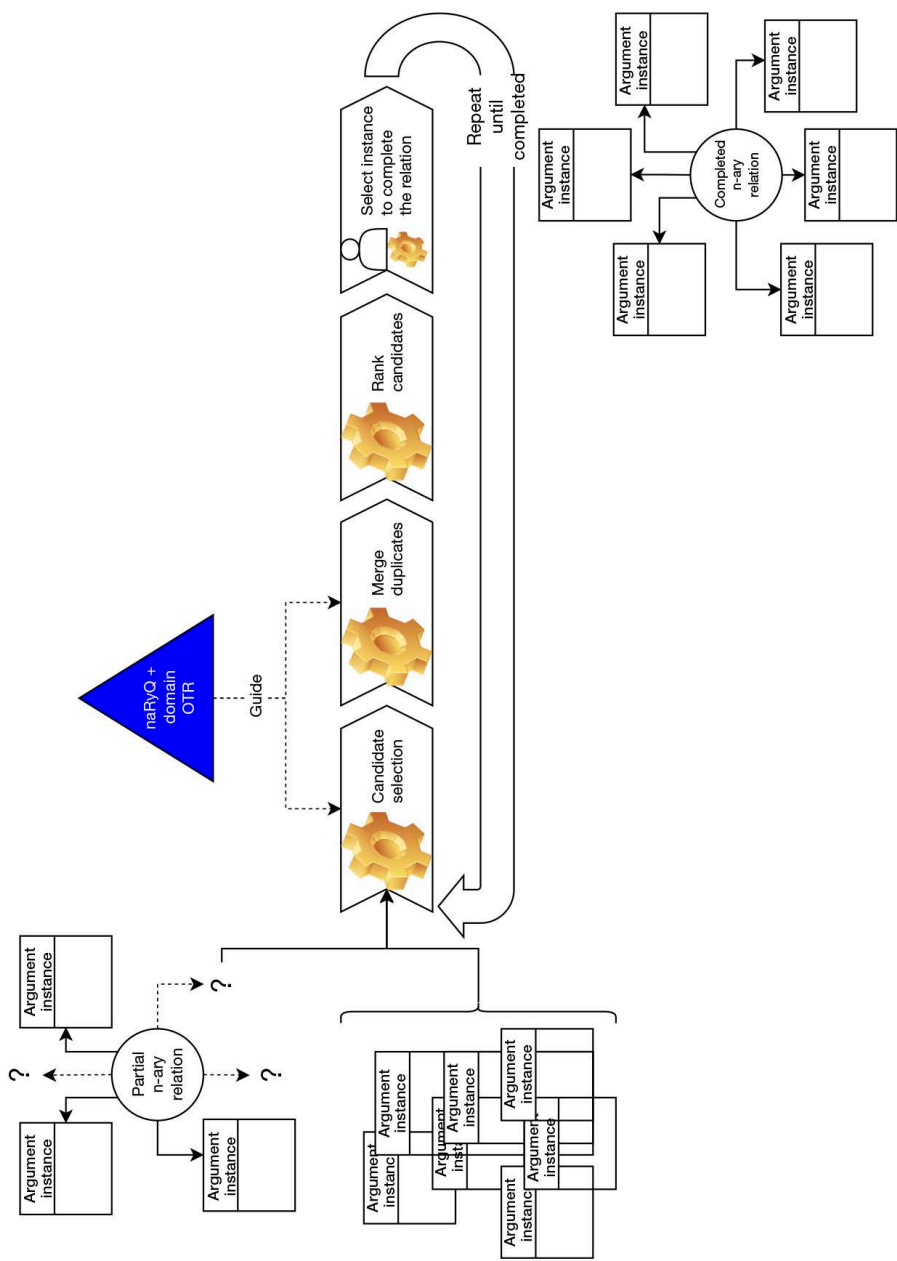


Figure 3.4. Reconstitution of n-ary relation instances.

A table and a STaRe representation obtained from it are illustrated in Examples 3.4 and 3.5. Since STaRe represents information from the table content, it does not contain the temperature or relative humidity values appearing in the caption. These argument instances were extracted during Phase I and will be reintegrated to the partial n-ary relation instance in the next step. The caption also mentions “*edible films*,” which is an instance of the packaging argument, but not a relevant one (too unspecific), and it should be filtered out.

Example 3.4. Table 4 from La Fuente Arias et al. (2022).

Thickness, Tensile strength (*TS*), Elongation at break (*E*), Elastic Modulus (*EM*), Water Vapour Permeability (*WVP*), Water Vapour Transmission Rate (*WVTR*), Oxygen Permeability (*OP*) and Oxygen Permeability Transmission Rate (*OPTR*) for monolayer and bilayer films with PLA.

Sample	Thickness (μm)	TS (MPa)	E (%)	EM (MPa)	WVP ($\text{g}\cdot\text{mm}/\text{kPa}\cdot\text{h}\cdot\text{m}^2$)	WVTR ($\text{g}/\text{m}^2\cdot\text{day}$)	OP x 10 ¹² ($\text{cm}^3/\text{m}\cdot\text{s}\cdot\text{Pa}$)	OTR ($\text{cm}^3/\text{m}^2\cdot\text{day}$)
Monolayers:								
Non-modified Starch	185 $\pm 20^a$	6.0 $\pm 1.0^e$	20.0 $\pm 3.0^a$	147 $\pm 30^c$	6.00 $\pm 1.00^a$	1453 $\pm 166^a$	0.092 $\pm 0.003^c$	4 $\pm 1.0^c$
DH Starch	200 $\pm 20^a$	6.4 $\pm 0.4^e$	23.0 $\pm 4.0^a$	193 $\pm 30^c$	0.30 $\pm 0.01^b$	48 $\pm 1.0^b$	0.086 $\pm 0.007^c$	4 $\pm 0.1^c$
PHBV_0	140 $\pm 1^b$	32.0 $\pm 3.0^b$	2.0 $\pm 0.2^b$	2013 $\pm 130^a$	0.04 $\pm 0.01^b$	11 $\pm 2.0^b$	0.160 $\pm 0.001^c$	11 $\pm 0.5^c$
PHBV_10	133 $\pm 2^b$	21.0 $\pm 1.0^c$	2.0 $\pm 0.2^b$	1412 $\pm 120^b$	0.23 $\pm 0.01^b$	71 $\pm 4.0^b$	1.800 $\pm 0.500^b$	115 $\pm 30^b$
PHBV_20	123 $\pm 7^b$	16.0 $\pm 2.0^d$	1.5 $\pm 0.2^b$	1306 $\pm 120^b$	0.21 $\pm 0.05^b$	59 $\pm 16^b$	4.400 $\pm 0.600^a$	306 $\pm 54^a$
PLA	133 $\pm 2^b$	37.0 $\pm 1.0^a$	3.0 $\pm 0.6^b$	2133 $\pm 90^a$	1. $\pm 0.03^b$	39 $\pm 5.0^b$	1.800 $\pm 0.100^b$	121 $\pm 8.0^b$
Bilayers with PLA:								
Non-modified Starch	323 $\pm 10^a$	17 $\pm 1^d$	5.0 $\pm 1.0^a$	1160 $\pm 70^e$	0.88 $\pm 0.20^b$	99 $\pm 27^{bc}$	0.229 $\pm 0.008^c$	6.0 $\pm 0.2^c$
DH Starch	301 $\pm 11^a$	19 $\pm 1^d$	5.0 $\pm 1.0^a$	1327 $\pm 70^d$	0.44 $\pm 0.04^c$	46 $\pm 4.0^c$	0.230 $\pm 0.006^c$	7.0 $\pm 0.1^c$
PHBV_0	253 $\pm 6.0^b$	42 $\pm 2^a$	3.0 $\pm 0.2^b$	1852 $\pm 60^a$	0.65 $\pm 0.20^{bc}$	96 $\pm 30^{bc}$	0.330 $\pm 0.050^c$	13 $\pm 2.0^b$
PHBV_10	240 $\pm 4.0^c$	36 $\pm 1^b$	3.0 $\pm 0.2^b$	1740 $\pm 20^b$	0.82 $\pm 0.15^b$	126 $\pm 23^b$	0.460 $\pm 0.040^b$	18 $\pm 2.0^b$
PHBV_20	243 $\pm 6.0^c$	31 $\pm 1^c$	2.2 $\pm 0.2^c$	1622 $\pm 30^c$	2.02 $\pm 0.01^a$	347 $\pm 25^a$	0.706 $\pm 0.002^a$	30 $\pm 2.0^a$

Means with different lowercase letters in the same column are significantly different ($p<0.05$)

Example 3.5. STaRe representation of the oxygen permeability partial n-ary relation instance present in the third line of the Example 3.4 table.

Ontological	Relation	
	Result Argument	Target
		Oxygen_Permeability
		Node
		Oxygen_Permeability
	Attached Value	Original_Value
		"0.160±0.001"
		"10 ¹² (cm3/m.s.Pa)"
		"OP"
	Arguments	Target
		Packaging
		Node
		PHBV
		Original_Value
		"PHBV_O"
		Attached Value
		"PHBV_O"
		...
		Target
		Thickness
		Node
		Thickness
		Original_Value
		"140±1"
		"µm"
		Attached Value
		"Thickness"
Structural	Table	Table 4
	Caption	"Thickness, Tensile strength (TS), Elongation at break (E), Elastic Modulus (EM), Water Vapour Permeability (WVP), Water Vapour Transmission Rate (WVTR), Oxygen Permeability (OP) and Oxygen Permeability Transmission Rate (OPTR) for monolayer and bilayer films with PLA."
	Segment	"Results"
	Document	https://doi.org/10.21203/rs.3.rs-2070061/v1

Reconstitution of n-ary relations

Using STaRe, SciPuRe, and the OTR, Phase II proceeds in three steps (see Figure 3.4). First, argument instances that might complement the missing arguments in the relation are selected, and duplicates are then merged. Finally, representation features are exploited to determine the best candidates. These steps are repeated until the relation instance is complete or no more valid candidates of argument instances can be found.

- **Candidate selection:** for a given missing argument in a partial relation, all possible candidates are selected thanks to the Argument, Target, and Document features of STaRe.
- **Duplicate merging:** in a given set of candidates, some argument instances may be duplicates of one another (i.e., the same information is repeated in different places in a document). Duplicates of a symbolic argument are determined using the Node feature (i.e., same specific OTR concept). Quantitative argument duplicates are determined using the Original_Value feature (i.e., same numerical value).
- **Candidate ranking:** two endogenous (i.e., dependent on internal resources) methods and a hybrid endogenous and exogenous (i.e., dependent on an external

resources) method can be used to choose the best candidates. All methods aim to measure the association between a given textual argument instance and the argument instances of a partial relation.

- The Structural method is based on the specificities that exist in scientific article structures (Cohen et al., 2010; Hofmann et al., 2009; Kando, 1997; Shah et al., 2003). The first way to proceed is to search for the closest possible context around the table that contains at least one candidate (i.e., caption, then subsection, then section, then document). Candidates present at the same level are discriminated using a token distance. A variation uses a guide relying on expert knowledge of associations between argument types and document sections (e.g., the Method argument should be looked for in the Materials and Methods section, then in Experiments, then in the Introduction).
- The Frequentist method is based on measuring co-occurrences (Cafarella et al., 2005; Ru et al., 2018; Sun, 2009; Yates et al., 2007) between a candidate and the argument manifestations of a partial relation (i.e., the textual instances of the table instances). These co-occurrences can be measured using association scores such as Jaccard (1901), Dice (1945), or pointwise mutual information (PMI) (Church and Hanks, 1990) and in different context sizes (e.g., section, subsection, sentence).
- The Word Embeddings method relies on the ability of pretrained language models to reflect similarity between two tokens (e.g., using the cosine value between their two vectors). This similarity can be used to measure the association between a candidate and the partial relation arguments. Since our argument instances `Original_Value` and `Attached_Value` include more than one word, we can consider the highest value or average them.

Our experiments and evaluations of different scores and their variations can be found in Section 3.3.4.

3.2.4 The atweb software tool

A web application was designed to apply the above method. It was created in Python using the Flask library³⁶ and can be used to assist the manual extraction of n-ary relations of gas permeabilities in scientific publications. This tool accepts scientific publications obtained from ScienceDirect.com in HTML format. It can be used as an assistant in the work of an annotator wanting to create data to be added to atweb. Once a paper has been uploaded, the tool offers different functionalities.

First, it performs the automatic annotation of argument instances in the full text. Uploading an article (Figure 3.5) automatically triggers the recognition of argument instances from the text. The instances are represented in the SciPuRe format and their respective relevance scores are immediately calculated. Uploading an article will also trigger the recognition of tables and automatic annotation of partial n-ary relations in tables. This will result in the creation of STaRe representations for the partial n-ary relations.

The annotator can then verify and manually correct the recognized argument and relation instances (Figure 3.6). The tool starts with the document tables and can modify the annotation of columns and cells. Columns indicate their related generic concept (e.g., Packaging, Temperature), along with the unit of measurement for quantity arguments.

36. <https://flask.palletsprojects.com/en/stable/>

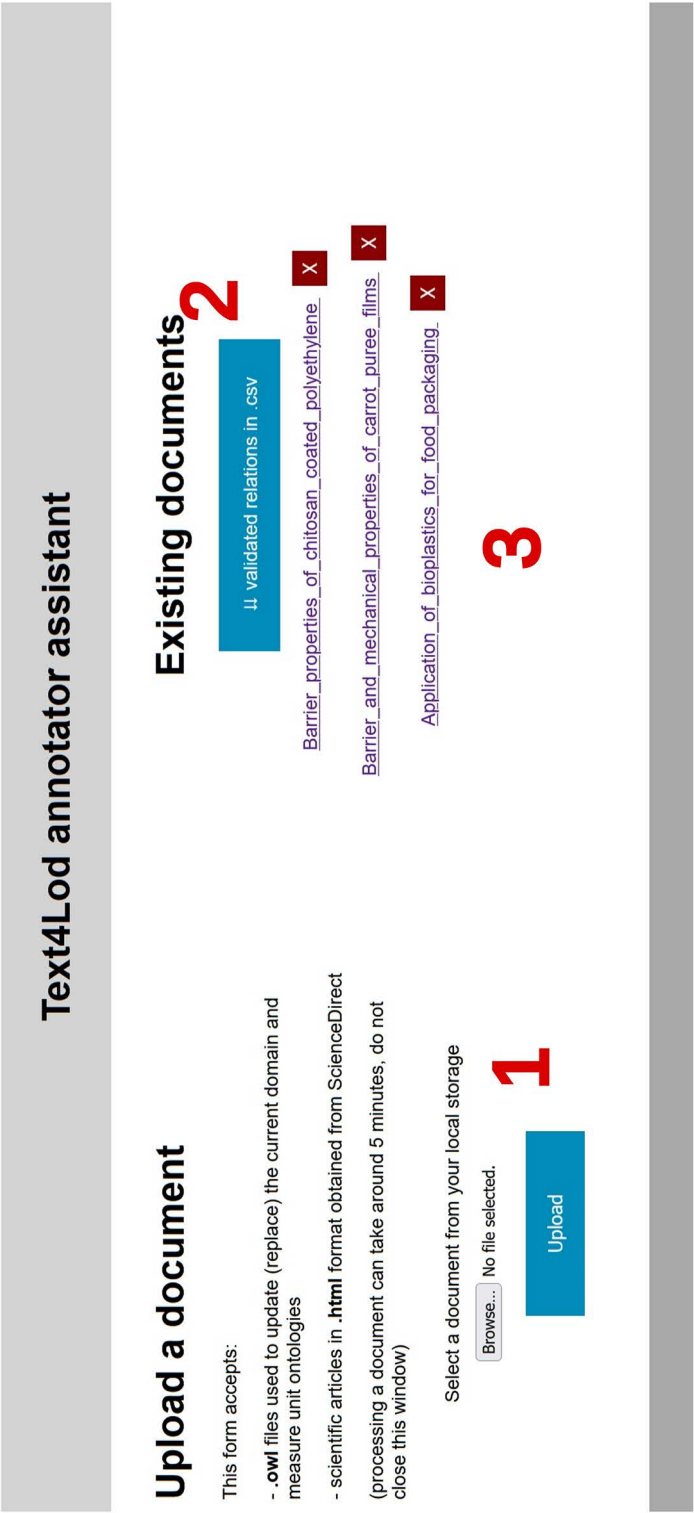


Figure 3.5. Home page of the Text4Lod annotation tool showing the three-step process: 1. Upload a document and wait until processing is complete; 2. Download a .csv file of all relation instances that have been manually validated; 3. Select a document to view and work on the annotations.

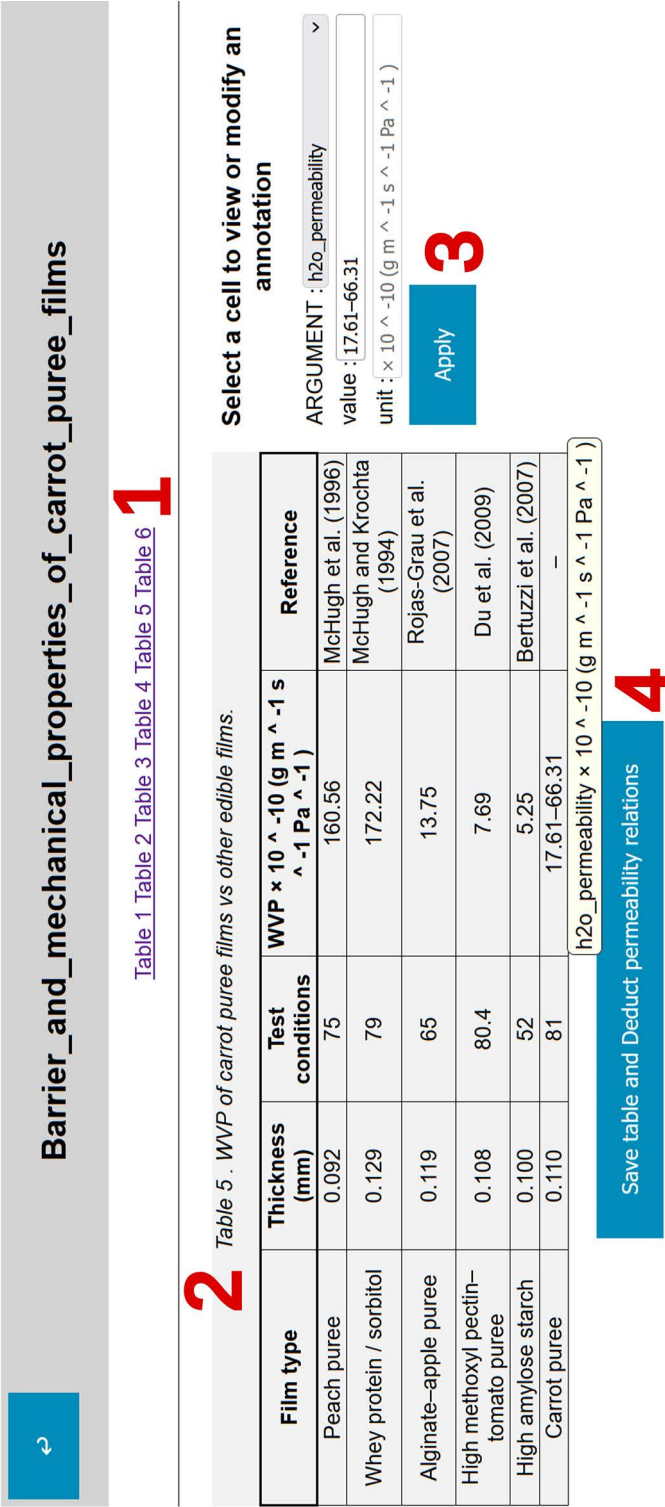


Figure 3.6. Four-step process to correct and validate partial n-ary relations: 1. Select the table to view; 2. Click a column or a cell to edit it; 3. Manually correct selected argument instance; 4. Validate the table and display the permeability relations (see Figure 3.7).

Cells contain the relation argument instances and can be selected to correct the value of a quantity argument or the specific concept of a symbolic argument. Once the tables have been validated, the partial n-ary relations can be retrieved and sent to the next step.

Argument instances can then be manually selected to complement partial n-ary relations (Figure 3.7). The user can fill in any argument instances that are missing in the partial relations, either manually or by selecting a suggested candidate argument. When clicking on a missing argument instance, the tool will automatically suggest the best candidates found in the text based on different methods. The annotator can then select a candidate argument instance, modify it if necessary, and add it to one or all relations.

Once an n-ary relation instance has been validated, it can be downloaded in a SPARQL format to be added to atweb. The home page also offers a way to download a .csv file³⁷ containing all relation instances that have been previously validated.

►► 3.3 Outcomes

Our pipeline to extract n-ary relations contains several steps, with the results of one influencing the next. Independent evaluations were conducted on Gold Standards of textual argument instances (Lentschat et al., 2021a, 2021b), partial relations in tables (Lentschat et al., 2021c, 2022c), and reconstituted n-ary relations (Lentschat, 2021). A full evaluation of the two-phase extraction of n-ary relations pipeline was also performed (Lentschat et al., 2022a). In this approach, we consider different levels of expert involvement to select valid n-ary relation instances.

3.3.1 Extraction of partial relations from tables

In our use case, the automatic table annotation process was evaluated (Lentschat et al., 2022c) on a corpus of 334 n-ary relations from 31 tables found in 10 scientific articles (Lentschat et al., 2021c). The code used to preprocess the tables and automatically annotate them is available along the data.³⁸ The automatic annotation method is based on existing work (Hignette et al., 2009) and extended with preprocessing to deal with columns and rows split into multiple ones, propagate information present in the header that should be in cells (e.g., typically units of measurement), and “rotate” the table if the first column functions as the header.

This method then assigns tags to the columns and cells. This first step concerns the recognition of symbolic and quantitative concepts found in the header as well as relation concepts. Relation instances present in rows are then recognized, and argument instances are attributed to cells using the information previously recognized in the header.

The tags assigned during the automatic annotation were compared to tags from a manual annotation in terms of recall, precision, and F-score. The results are consistent and comparable with the original study (Buche et al., 2013a; Hignette et al., 2009). The recall scores of all tags are high (i.e., their average is 0.77). The precision on symbolic concepts tags is low (i.e., 0.38) due to the large number of columns containing a term relating to a symbolic concept that does not describe the content of the column.

37. https://gitlab.softwareheritage.org/eskode/webapp_ARTEXT4LOD/-/raw/master/demo_webapp_TEXT4LOD.mp4?ref_type=heads

38. <https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/GCZBC9&version=3.0>

Relations detected

1

O2_Permiability	Packaging	Temperature	Method	Partial_Pressure	Thickness	Relative_Humidity	VALIDATE
[1.04 ± 0.15 ^a × 10 ⁻¹⁷ g / m s Pa]	[Aqueous acid solvent ^a , 'chitosan(aqueous acid solvent)']	[25 ^a , °C]	?	?	?	?	↕
[8.76 ± 2.58 ^a × 10 ⁻¹⁷ g / m s Pa]	[Aqueous acid solvent ^a , 'chitosan(aqueous acid solvent)']	[25 ^a , °C]	?	?	?	?	✓
[2.22 ± 0.27 ^a × 10 ⁻¹⁷ g / m s Pa]	[Hydroalcoholic acid solvent ^a , 'chitosan(hydroalcoholic acid solvent)']	[25 ^a , °C]	?	?	?	?	✓
[4.98 ± 1.10 ^a × 10 ⁻¹⁷ g / m s Pa]	[Hydroalcoholic acid solvent ^a , 'chitosan(hydroalcoholic acid solvent)']	[25 ^a , °C]	?	?	?	?	✓

Max number of candidates to retrieve : 5 in each method

2

Select an argument to complete

ARGUMENT : Temperature

concept : Temperature

valeur : 25

unit : °C

☒ add to every relations

ADD

Suggestions : 9 / 50

Combined method (top results appearing in both methods)

1) 'temperature', [2], [°C]

Materials and methods - Materials and reagents :
"Anhydrous glycerol (GLY) (Fluka Chemical , 98 % purity , Germany) was used as a plasticizer in order to improve the mechanical film properties. Acetic acid (glacial 100 % , Merck , Darmstadt , Germany) and pure ethanol (absolute , Sigma-Aldrich) were used as solvents in the preparation of the film forming solutions (FFS) . Magnesium chloride (MgCl₂ , Sigma) , sodium chloride (NaCl , Sigma) and potassium dichromate (K₂Cr₂O₇ , Sigma) were used to prepare saturated salt solutions to fix the relative humidity for water vapour and gas permeability measurements ."

Semantic method (context similarity with existing arguments)

1) 'temperature', [2], [°C]

Results and discussion - Oxygen, carbon dioxide and air permeability - Influence of casting solvent and glycerol on gas permeability of chitosan self standing films - table :
"Oxygen, carbon dioxide, air permeability and CO₂/O₂ permselectivity at 25 °C and 0 % (dry gas) and > 96 % (humid gas) of chitosan films influenced by casting solvent and glycerol addition in film formulation ."

Figure 3.7. The three-step process for assisted completion of partial n-ary relations: 1. Select a missing argument to display suggested candidates (here, the user set the maximum number of candidates to 5 per method, 9 were actually found); 2. Enter manually or modify a candidate, then add it to a relation; 3. Click a suggested candidate to modify or validate before addition.

These false positives present a small risk of error propagation, as these columns rarely contain instances of arguments. Errors in the annotation of quantitative concept tags (i.e., $p = 0.62$) mainly stem from the recognition of units of measurement. The annotation of relation instance tags is satisfactory, especially as it relies mainly on the identification of the previous tags (i.e., $p = 0.67$). Similarly, the annotation of relation instances and argument instance tags relies on the correct annotation of the previous tags, so their results are similar (i.e., $p = 0.69$ and 0.66).

3.3.2 Argument instance extraction evaluation

To evaluate the extraction process, we used a Gold Standard³⁹ of 1,772 argument instances (Lentschat et al., 2021a, 2021b). This Gold Standard comprised 50 open access articles obtained in HTML format from the ScienceDirect website. It was annotated by three annotators on a WebAnno server (Eckart de Castilho et al., 2016) and they obtained a Gwet’s Kappa (Gwet, 2014) value of 0.62.

We evaluated the performance of our argument instance extraction approach and the effects of the preprocessing put in place using recall, precision, and F-score values. Results are displayed in Table 3.1; detailed evaluations for each argument type can be found in previous publications (Lentschat et al., 2020, 2022b).

Table 3.1 Overall results of argument instance extraction

Approach	Recall	Precision	F-score	No. extracted
Baseline	0.74	0.38	0.50	14,603
+Terminological variations	0.80	0.37	0.51	18,749
+Acronyms	0.79	0.36	0.49	15,664
GLOBAL	0.85	0.41	0.55	20,669

Table 3.1 shows that recognizing terminological variations and acronyms increases the coverage of OTR and positively affects the argument instance extraction (i.e., recall rises from 0.74 to 0.85). Using both also slightly increases precision, which underlines the importance of combining both processes and shows that the terminological variations and acronyms added to the resource are relevant to the task.

The precision scores highlight the need to filter out false positives. Nonrelevant argument instances are extracted for different reasons depending on their types. Most false positives of symbolic arguments come from terms present in the OTR vocabulary that are not specific enough to be considered valuable instances (e.g., the term “packaging” is very frequent but is not a valid instance of the Packaging argument). Quantitative arguments are subjected to interfering information (e.g., temperature values present in the texts but unrelated to permeability relations), which also creates a major source of false positives.

Overall, the extraction of argument instances shows a satisfactory recall score, proving that most of the information sought is extracted, but an underwhelming precision score prompts the need to select only information of interest. In the next subsection, we will present different relevance scores and evaluate their ability to filter out false positives.

39. <https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/U7HK8J>

3.3.3 Relevance score evaluation

To evaluate the designed relevance scores, we used the Precision@N (Craswell, 2009) measure (also called Precision@K [Manning et al., 2008]). This score allows us to evaluate, for a given relevance score, the precision value for a set of the top-N ranked argument instances. This enables us to manually determine thresholds to filter out false positive argument instances according to their relevance score. Examples of Precision@N are shown in Figure 3.8.

An analysis of Figure 3.8. provides several insights. First, it appears that the semantic relevance score (CD) effectively ranks symbolic argument instances and that no true positives can be found after a certain point (i.e., the monotonic drop seen in Figure 3.8a). Concerning lexical relevance scores in Figure 3.8b, a term frequency (TF) score relying on the SciPuRe lexical and Document features also allows false positives to be sorted out from symbolic argument instances.

Our detailed experiments (Lentschat et al., 2020, 2022b) showed that the two scores used in Figure 3.8 are best suited for symbolic arguments and can be combined for enhanced results (Figure 3.9). Quantitative argument instances are best served by an ICF score, using the SciPuRe features of Attached_Value (i.e., denoting the quantity concept) and Segment (i.e., the type of section it appears in). This highlights the importance of taking document structures into account to describe expert knowledge from scientific publications.

Combining relevance scores (e.g., Figure 3.9) leverages the many SciPuRe features. Our experiments showed that linear combination produced no real gains, but that sequential combination offered good applicability to symbolic arguments. Ranking instances using the CD semantic score made it possible to prioritize specific information and sort out the most frequent false positives (e.g., the term “packaging”) before applying a lexical relevance score that could be influenced by term frequency.

3.3.4 Relation reconstitution evaluation

Our relation reconstitution approach was evaluated using a Gold Standard of n -ary relations representing gas permeabilities and composition relations.⁴⁰ This Gold Standard is based on partial n -ary relations manually annotated in tables (see Section 3.3.2) and was then manually completed with the argument instances previously annotated in texts (see Section 3.3.3).

The results of our n -ary relation reconstitution approach were evaluated using precision, recall, and F-score values measured based on the number of argument instances effectively missing from the partial relation (i.e., some relations were incomplete since all information was not systematically present in the document). The methods for selecting candidate argument instances, presented in Section 3.2.3, were designed to be used jointly with expert input. A set of n -best candidates (1, 3, 5, or 10) was thus used and evaluated. We also applied filtering using relevance scores computed at the end of Phase I.

The detailed results and analysis of our evaluation can be found in a previous publication (Lentschat et al., 2022a). Our main conclusions are that filtering a large number

40. <https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/1BBJBQ>

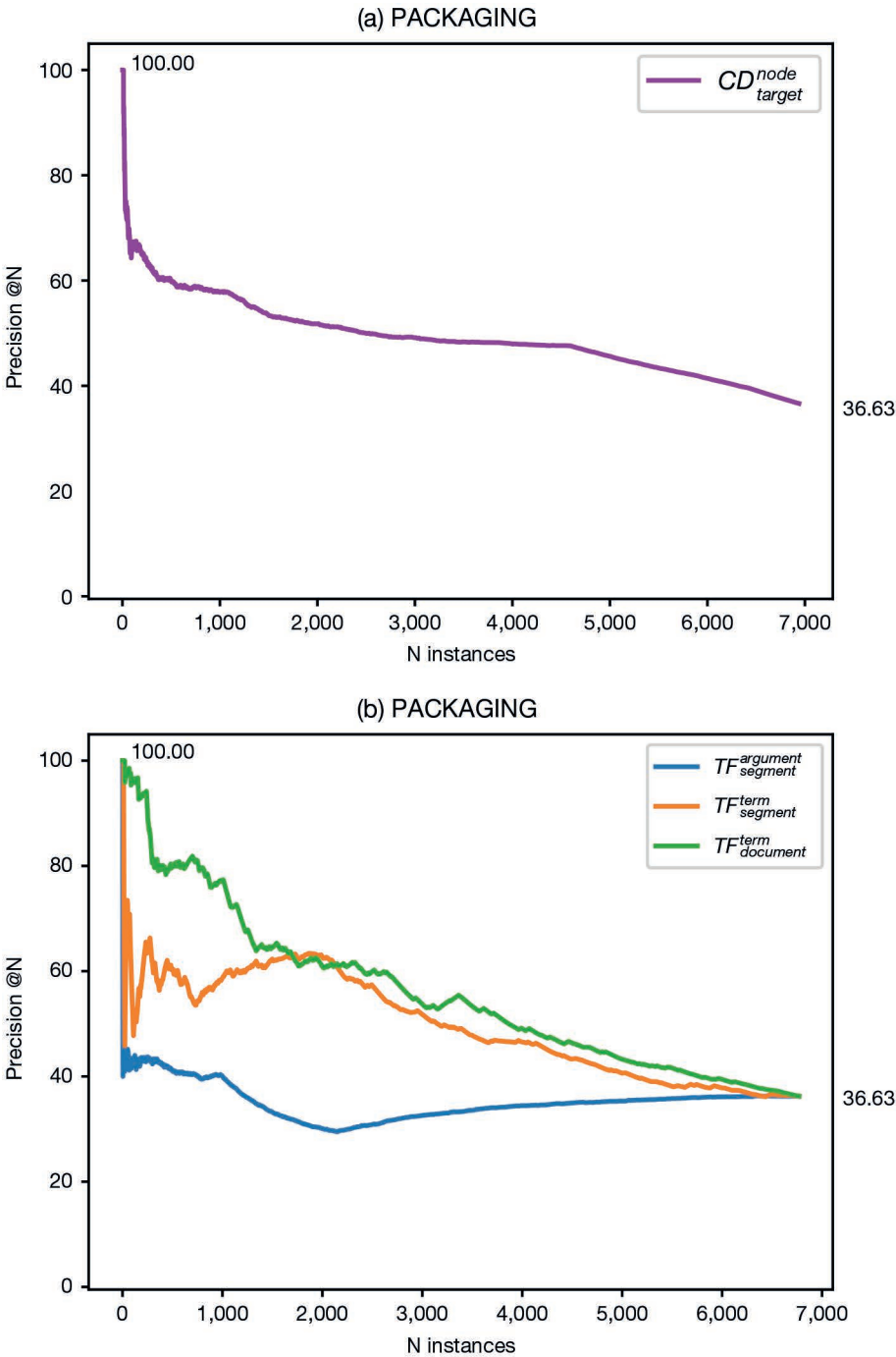


Figure 3.8. Precision@N evaluation of term frequency (TF) and CD scores on extracted Packaging argument instances.

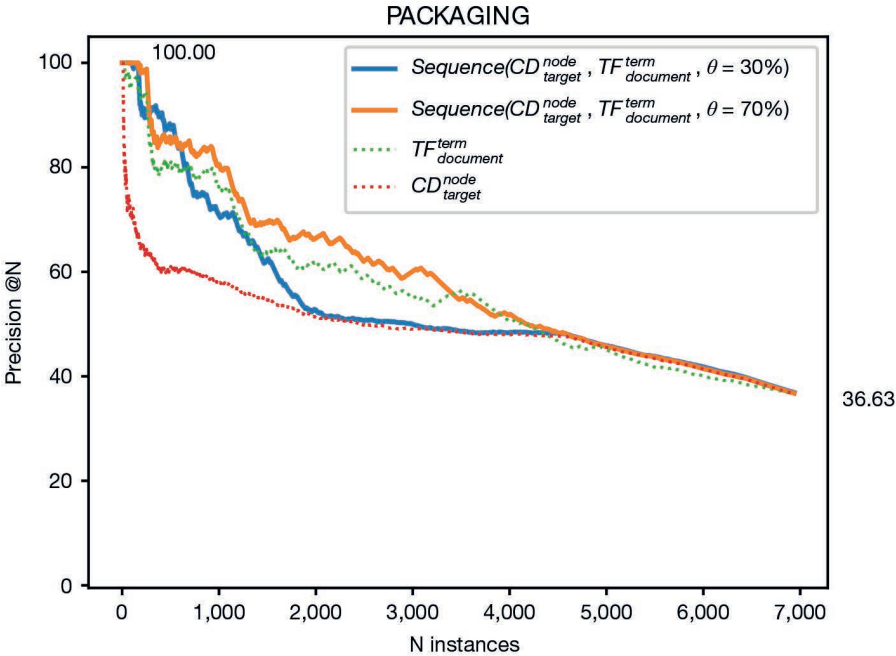


Figure 3.9. Sequential combination of CD and term frequency (TF) scores on extracted Packaging instances.

of argument instances at the end of Phase I carries the risk of discarding true positives and losing information. The proportion of instances filtered out can be adapted to the argument types (e.g., Packaging instances present more false positives than Method instances) using expert knowledge of the domain, but overall, a filtration of 20% produced the best results. The three methods used – Structural, Frequentist, and Word Embeddings – should be chosen depending on the desired level of human involvement.

Table 3.2 shows the F-score values for the best configuration found. The Structural method was the best at resisting noise due to the selection of a higher number of candidates, thanks to the manual guiding via the Argument-Sections. As such, this

Table 3.2 Best F-scores obtained (with 20% of instances filtered by relevance). PMI: point-wise mutual information

Method	Criteria	Number of candidates			
		1	3	5	10
Structural	Simple	0.35	0.58	0.58	0.65
Structural	Guided	0.45	0.56	0.61	0.74
Frequentist	Jaccard	0.48	0.54	0.61	0.66
Frequentist	Dice	0.46	0.55	0.60	0.66
Frequentist	PMI	0.44	0.53	0.60	0.68
Word Embeddings	core_web_trf	0.40	0.59	0.64	0.70
Word Embeddings	core_sci_scibert	0.39	0.57	0.65	0.70

method is best suited to performing relations that are later assigned to an expert. In an automated process, the Frequentist method would perform best, using a Dice score on instance manifestations (i.e., co-occurrences of Attached_Value SciPuRe features at a Document level). The results of the Word Embeddings method fall somewhere in between the two, requiring some human work to select the right candidates. It is no surprise that language models using a transformer architecture provide the most insights on which candidate to select (although one might argue that measuring semantic similarity between, for example, a Packaging and a Temperature argument is a stretch, and that the use of language models in our use case is another way of measuring co-occurrences).

► 3.4 Conclusion and resources

New solutions are needed to support the retrieval of scientific knowledge from publications. They would have several highly useful applications: help navigate the ever-increasing number of publications, assist the conception and writing of meta-analyses on a specific research question, and provide knowledge to expert decision systems. Articles that are publicly available on online platforms (e.g., Science-Direct) contain information that needs to be extracted and processed (i.e., described, structured, or represented) in such a way as to be actionable by downstream software.

In this chapter, we presented a way to semi-automatically feed a knowledge base with information extracted from a corpus of scientific articles. To this end, we used an ontological and terminological resource (OTR), created as a specialization of an nRyQ task ontology presented in Chapter 2 (TRANSMAT). Our application domain was food packaging characteristics, specifically gas permeabilities, and the TRANSMAT OTR guided the extraction and representation process. This resource formalizes knowledge as n-ary relations and is structured around two main parts:

- A task ontology representing the concept of the n-ary relation itself, the arguments that compose those relations, and concepts common to experimental domains (e.g., units of measurement, quantitative or symbolic arguments).
- A domain ontology that describes each domain-specific concept. For example, TRANSMAT describes an Oxygen_Permeability_Relation as the Oxygen_Permeability measurement of a Packaging, using a Method and under specific experimental conditions (e.g., Temperature, Thickness). Each domain concept comes with sub-(sub-) concepts and a vocabulary that contains alternative writing that is useful in identifying its instances in a text.

The main approach presented in this chapter is guided by the OTR. It is thought of as a generic way to extract n-ary relations from texts, with only the domain part of the OTR needing to be changed.

The information to extract from the scientific articles is given in the full-text parts of the documents as well as in the tables. We designed a two-phase method, a classical approach in the literature on n-ary relations (Berrahou et al., 2017; Zhou et al., 2014):

- **Phase I** entails extracting the relations argument instances found in the text. Guided by the domain OTR structure and vocabulary, the process involves first searching for new units of measurement, terminological variations, and acronyms to expand the vocabulary. This is done to not miss relevant information that would be required later. Argument instances are then extracted along with a representation,

SciPuRe, containing ontological, lexical, and structural features. Since most instances of arguments extracted are not related to gas permeabilities (i.e., the knowledge we are interested in), their relevance is evaluated at the end of Phase I. SciPuRe rich features allowed us to design relevance scores adapted to different argument natures (e.g., symbolic or quantitative) and able to filter out many of the false positives.

– **Phase II** is designed to reconstitute n-ary relation instances by completing partial relations found in an article table with argument instances extracted during Phase I. We employed an existing method (Hignette et al., 2009) to extract information from the tables as n-ary relations. These relations are partial (i.e., missing some argument instances), and represented in STaRe. The process to rebuild n-ary relations relies on STaRe and SciPuRe features, using them to find the missing argument instances among those filtered in Phase I and then to rank candidates.

Our experiments showed that SciPuRe and STaRe features are essential to reconstitute n-ary relations. First, the many argument instances extracted during Phase I contain a significant proportion of false positives (e.g., results not related to the n-ary relation of interest or too unspecific to be relevant). An efficient ranking of argument instances depends mainly on their nature. Scores measuring the association between argument appearance in the text and the document structure (i.e., lexical and structural features) are better suited to quantitative arguments. Symbolic arguments must be ranked in two steps: unspecific instances must first be discarded using ontological features, and then the most frequent instances identified using lexical and structural features. We observed that an overall filtration of 20% of the least relevant argument instances had a mostly positive impact, filtering out a large part of false positives without discarding many true positives. This level of filtering can be adapted to the argument type. For instance, the set of Packaging instances found in the text contained a larger number of false positives (i.e., since the word “packaging” is recognized but is not relevant as an argument instance) than the Method instances. Likewise, a variety of Temperature instances were recognized, but most of them did not relate to permeability measurements while Thickness values were nearly always related to packaging.

The two-phase method presented in this chapter can have different outputs depending on the desired level of reliance on experts. The experiment conducted on Phase II showed that different approaches to reconstituting n-ary relations yielded different results and suggested different strategies. One strategy was to have a fully automated process by using the frequent associations between argument occurrences. However, our results showed that this does not yet produce satisfactory results and could not be used by downstream processes. We recommend keeping humans in the loop. Relying on humans to feed information into a knowledge base is costly in terms of both time and money because the task must be performed by domain experts. Our method can make their work faster and easier. The tool presented in Section 3.3.5 will present potential n-ary relations to an expert that they can then validate. They also can correct the relation, browse candidate argument instances, or fill the information manually.

Designing tools to facilitate the management of scientific knowledge is an important undertaking today, and it concerns both researchers in the public sector and private professionals. We are convinced that keeping humans in the loop is an essential aspect of handling expert knowledge. This ensures that mistakes can be corrected, missing information can be completed, and the process remains responsible and reflective.

Recent developments in AI and NLP have seen the rise of large language models. Based on a transformer architecture, these models are trained on ever-larger text corpora to follow textual instructions (Ouyang et al., 2022). These models are then tailored to be used in a specific context, such as providing information in a chat and answering questions, or being used as an agent to perform specific tasks.

Although an LLM's performance in a specialized domain depends in part on its training data (niche domains remain a challenge), current research shows that LLMs could be used at different stages:

- Ontology construction: LLMs can extract the main concepts and draft the taxonomy of a specific domain. This could assist experts in building domain ontologies (Kommineni et al., 2024) upstream of the extraction process.

- Argument instance extraction: Named Entity Recognition (NER) approaches using LLMs (Wang et al., 2023) can be leveraged to extract argument instances. Using models trained to follow instructions and a prompting approach based on few-shot examples is essential to obtain decent results.

- Argument instance filtering: one could consider providing description features (e.g., SciPuRe) to an instruction model to measure the relevance of extracted argument instances. However, to our knowledge there has been no research on a comparable task and this would require further investigation and experiments.

- N-ary relation reconstruction: relation extraction relying on LLMs has largely been studied in the bioclinical field. The main challenge at this time remains the specificities of other specialized fields.

- Expert system and question answering: downstream of the knowledge base instantiation, an LLM can replace the software interface of an expert system. Using approaches such as retrieval-augmented generation (Gao et al., 2024), the model can search the knowledge base to answer the user query with accurate information.

Available resources

Available resources					
Research question	Proposed solution	Main scientific papers	Data papers and datasets	Software codes	Courses and technical reports
Q1	ARTEXT4LOD method to extract n-ary relation instances guided by an ontological and terminological resource (OTR).	OTR-guided data table annotation (https://hal-lirmm.ccsd.cnrs.fr/lirmm-00642899v1) OTR-guided textual argument extraction (https://hal.inrae.fr/hal-03616243v1) Representation of n-ary relation arguments (https://dl.acm.org/doi/10.1145/3405962.3405978) OTR-guided extraction of n-ary relation instance (https://hal.science/hal-03767632v1)	Data paper on OTR-guided textual argument extraction (https://hal.science/hal-03250848v1) and its dataset (https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/U7HK8l) Data paper on OTR-guided extraction of table n-ary relation instances (https://hal.science/hal-03610433v1) and its dataset (https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/GCZBC9) Dataset of reconstituted n-ary relations (https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/1BBJBQ)	HAL: https://hal.science/view/index/docid/5209623 Software Heritage: https://gitlab.softwareheritage.org/eskodo/webapp-ARTEXT4LOD/-/raw/master/ https://gitlab.softwareheritage.org/eskodo/webapp-ARTEXT4LOD/mp4?ref_type=heads	https://gitlab.softwareheritage.org/eskodo/webapp-ARTEXT4LOD/-/raw/master/ https://gitlab.softwareheritage.org/eskodo/webapp-ARTEXT4LOD/mp4?ref_type=heads

Chapter 4

Reasoning About Process: Experimental Data and Technological Expert Know-how for Decision Support

*Mélanie Münch, Cédric Baudrit, Julien Cufi, Julien Couteaux, Christophe Fernandez,
Luc Menut, Alrick Oudot, Patrice Buche*

►► 4.1 Introduction

Agrifood chain processes rely on a wealth of knowledge, expertise, and experience accumulated over time. Companies that specialize in food product processing draw on their know-how to adjust their practices in response to raw material variations, consumer expectations, and regulatory changes. Gaining knowledge through hands-on experience is a common practice in this industry, enabling workers to develop significant expertise. This knowledge is typically passed down through on-the-job training and learning by doing. However, recent economic and health crises, along with internal challenges such as higher employee turnover and recruitment difficulties in certain sectors, have made it increasingly difficult to preserve and transmit this valuable know-how.

In the food and biobased products industry, a production process encompasses the steps through which raw materials (inputs) are transformed into one or more final products (outputs). Analyzing these processes is particularly challenging due to their heterogeneity (involving different variables and scales) and the time-dependent flow of information. Providing relevant conclusions and recommendations requires developing adequate modeling tools that can integrate, to the fullest extent possible, available knowledge that is heterogeneous in nature and quality. Such modeling tools must be able to manage heterogeneous sources of knowledge (experimental data and expert opinions), multiple manipulated scales, and different forms of uncertainty.

The shift to Industry 4.0 in food and biobased products application domains makes numerous numerical models available. For example, in the field of food packaging design, modified atmosphere packaging (MAP) has proven to be an efficient way to delay senescence and spoilage. To facilitate MAP design, mathematical models, known as virtual MAPs, have been developed by researchers working in this field to design passive (Mahajan et al., 2007; Sousa-Gallagher and Mahajan, 2013) or active MAP (Cagnon et al., 2013; Charles et al., 2003, 2005) for fresh and fresh-cut fruits and vegetables. Such numerical tools simplify the design steps by predicting relevant optimal recommendations. For instance, prediction of gas permeability values permits

the MAP to reach the recommended atmosphere for the target product and therefore to extend shelf life while maintaining the quality and safety of the packaged food (Munch et al., 2023b). However, all the aforementioned tools only considered some requirements, but numerous other requirements may come into play and affect the final decision. These requirements may be related to processing, marketing, commercial, or distribution concerns (e.g., cost, processability of inputs, industrial feasibility, environmental impact, safety, waste management, etc.). Aside from the constraints of some requirements (e.g., food shelf life and quality, correlated to the material's gas permeation rates), packaging also relies on user preferences, naturally expressed as wishes (e.g., transparent material would be preferred) or constraints (e.g., cost of raw material must be less than €3/kg). To help food-chain stakeholders choose a relevant solution compared to others (e.g., a packaging material that would meet all the requirements of a given fresh fruit or vegetable), multicriteria decision support tools based on numerical models and collection of possible solutions stored in a database need to be developed. Such tools must take into account that the collection of possible solutions may be incomplete in the sense that some of the characteristics associated with solutions and corresponding to requirements may be unknown.

The research questions addressed in this chapter are:

- Q1. How can technological recommendations to solve a production defect or reach a given quality be determined using collective know-how expressed as qualitative knowledge?
- Q2. How can the absence of a deterministic model be managed from a reverse-engineering optimization perspective?
- Q3. How can we extract information from an incomplete data source containing candidate solutions using multicriteria queries while taking into account constraints and wishes for decision-making purposes?

State of the art and positioning, including presentation of available scientific papers describing our proposal

The research questions are guided by the notion of expert knowledge and its diversity. The questions deal with three different levels of knowledge: qualitative only (Q1), quantitative but without models (Q2), and quantitative with models (Q3). Readers interested by the comparison of the state of the art in terms of expert knowledge representation are invited to go to Chapter 2 for complementary information about ontology-based knowledge representation models to solve this research question. In the following, the comparison is focused on the topic of this chapter: reasoning about process.

To address the first research question Q1, significant progress has been made in knowledge integration and multicriteria analysis methods and tools in the field of food science and technology. However, these advances remain fragmented and incomplete (Aceves Lara et al., 2018; Thomopoulos et al., 2019). Various methods have been developed to collect scientific and technological knowledge and data for different purposes, but they have generally been limited to elementary processing operations, failing to encompass entire processing chains.

For instance, Kansou et al. (2014) proposed a qualitative model for a unitary mixing operation, using an expert system to predict wheat flour dough quality. Baudrit et al.

(2015) modeled preharvest grape berry maturity – a crucial factor for the wine industry – by combining expert knowledge, data, and probabilistic graphical approaches. Belna et al. (2022) optimized microfiltration unit operations to balance conflicting stakeholder objectives, such as maximizing product quality while minimizing costs and environmental impacts. Similarly, Baudrit et al. (2022) compiled data from scientific articles to model the entire milk microfiltration process, incorporating multiple unit operations such as skimming, heat treatment, and storage in addition to the microfiltration step. Those data are available in Buche et al. (2021b). However, the method described in Baudrit et al. (2022) is limited to training a predictive model of the milk microfiltration unit operation under large-scale operating conditions, considering different membrane types.

In the context of a circular economy, Belaud et al. (2022) developed a decision support system to rank alternative lignocellulosic waste conversion processes, leveraging knowledge engineering tools to compile experimental data and assess potential environmental impacts. Munch et al. (2021, 2022c) combined ontologies and probabilistic models and linked open data to identify agricultural waste suitable for biocomposite production in food packaging through reverse engineering. In both cases, ontological models facilitated an analysis of the entire processing chain's impact on either the final product's quality or key process indicators. However, these approaches require extensive numerical experimental data to yield reliable results.

Given their high data requirements, these methods are unsuitable for achieving the objectives outlined in this research question, as they fail to represent the full spectrum of collective knowledge across an entire food chain.

To address the second research question Q2, machine learning approaches are used to make up for the lack of predictive models based on expert knowledge. Due to the vast variety of models, this chapter will not provide an exhaustive list of machine learning algorithms and their principles. Readers curious about this topic could refer to Freire et al. (2025), which presents an overview of the different applications in dairy processes.

In this section, machine learning models will be considered as generic models $f(X) = y$, where f is a trained algorithm, X a selection of feature variables, and y a predicted variable. During a training phase, the machine learning algorithm's parameters are tuned so that the obtained model can predict an output y with a confidence for a new given set of input data X . Data-driven models can make up for the lack of expert or mechanistic models by relying on statistical analysis to draw underlying relations between the variables within a raw data selection (quantitative, qualitative, or mixed) (Mavani et al., 2022).

For example, machine learning models have proven useful for modeling CO₂ solubility in modified atmosphere packaging (Munch et al., 2023b). While no general model existed, a model trained on a dataset of 362 food products was able to mimic the CO₂ solubility of a food product based on its composition and temperature; this new global model was then validated using expert knowledge in the form of partial models. Similarly, Perrignon et al. (2025) presented a way of modeling the industrial performance of a cheesemaking industry based on machine learning modeling, a dataset of 3,500 cheese processes, and expert knowledge for validation. It should be noted that the first example relied solely on the existing features present in the input X (the raw

values of food product composition and temperature), whereas the second model relied on feature engineering with pretreated parameters (the milk composition and process description were normalized).

Indeed, while a firsthand model theoretically relies solely on raw data, the main pitfall of data-driven models is their dependence on statistical observations. In particular, they cannot distinguish by themselves correlation from causation. On this topic, Spirtes et al. (2001) detailed a set of criteria describing data quality that can affect the overall model quality, such as missing data, selection bias, measurement error, or deterministic cases. More recent approaches have addressed other issues that might impair the data-driven modeling task, such as the presence of outliers, high dimensionality, noisiness, or imbalanced classes (in the specific case of classification tasks) (Corrales et al., 2018a, 2018b, 2020). This data reliability assessment is necessary before the training phase, as poor quality can lead to incorrect outcomes such as overlearning, which lends too much weight to nonrepresentative data points.

To complement the raw data, expert knowledge from the modeled domain can also be included. Indeed, combining data with expert knowledge can provide a new understanding of patterns in a dataset by guiding the possible pretreatments or even introducing new engineered feature variables (see Section 4.3.2 for an example of introducing new features).

As such, the need for expert knowledge to validate both the dataset and the machine learning models is paramount to (i) mitigate their effects and produce reliable results according to expert know-how and (ii) improve models by identifying knowledge gaps. To avoid the tendency of “black box” models (Vilone and Longo, 2021) and aid this task of combining learned models with expert knowledge, Munch et al. (2022c) proposed an approach based on a type of machine learning model, graphical probabilistic models, and an expert knowledge-driven structure using ontologies. The resulting causal model is then used in a reverse-engineering context to provide recommendations on biomass options for use in biocomposite processing. While these kinds of approaches help compensate for the lack of deterministic models, they still need to be paired with expert preference criteria to guide a decision-making process.

To address the third research question Q3, we will consider preexisting models and expert knowledge in the form of preference criteria. In this context, experts query existing databases containing candidate solutions using their preferences and results that are partially determined by simulation models.

The previous application presented the training of a model to simulate the process of different biocomposite materials (Munch et al., 2022a). When given a collection of candidate biomass solutions stored in a database and a set of expert preference criteria on the final material, this model was able to generate recommendations. This approach did not, however, allow the handling of (i) uncertainty in the database and (ii) user criteria expressing either constraints or wishes.

Uncertainty can have many causes. Destercke et al. (2011) introduced cases where values are tainted by biological variability or tolerance threshold indicated by users. Information can also be incomplete (missing values), preventing reasoning on the whole domain. As such, uncertainty adds complexity to the reasoning and thus to the generation of recommendations.

On the other hand, constraints and wishes are both possible expert preference criteria that help shape recommendation generation. They are bipolar expressions of negative and positive parts of the desired result (Dubois et al., 2013; Dubois and Prade, 2009). Sometimes these parts are of the same nature and can be combined into a single one (e.g., “ x must be greater than 5” and “ x must be less than 10” can be combined into “ x must be between 5 and 10”). On other occasions, however, these parts may be different and therefore cannot be processed as a single piece of information (“the cost must not be too great” and “the quality should be the best”). To this end, asymmetric bipolarity has been coined (Cacioppo et al., 1997) to characterize cases where expectations can be both positive and negative. In the case of database queries, it can distinguish between constraints (negative preferences) that must be satisfied and wishes (positive preferences) that should be respected. These make it possible to propose a solution ranking that best satisfies the expert.

To combine both database uncertainty and flexible preference criteria, much research has proposed the use of fuzzy sets, which have proven useful in handling flexible queries using their gradual scale (Bosc et al., 1994). In particular, Bosc and Pivert (1995) proposed an extension of the SQL language called SQLf to introduce fuzzy predicates processed on crisp information. Finally, other approaches have also been proposed to introduce preferences into queries in the database community (Bruno et al., 2002; Chomicki, 2003; Kießling and Köstler, 2002). However, these do not consider query bipolarity, whereas others do (De Tre et al., 2010; Dubois and Prade, 2009). In this chapter, two types of preferences will be covered, constraints and wishes. The querying mechanism considers that candidate solutions stored in the database may be incomplete. In case of an empty set of answers, a guaranteed set of results is proposed that has been determined as the closest solutions compared to requirements expressed by the user.

Figure 4.1 presents the structure of the following sections, with all research questions and their associated sections and prerequisites.

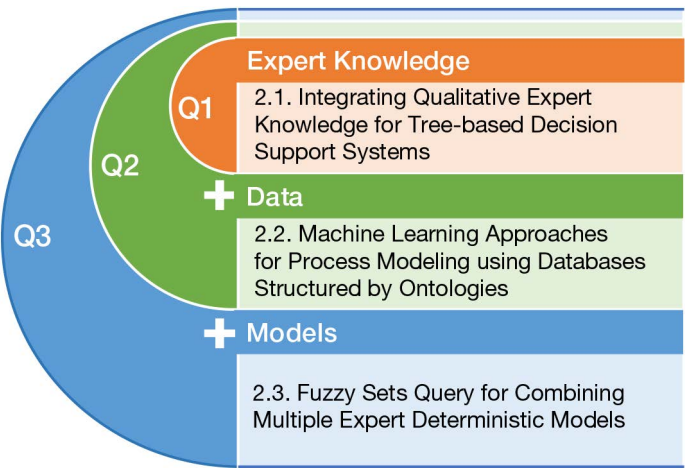


Figure 4.1. Structure of the chapter. Each step answers a research question and presupposes a different level of information: the first (2.1) requires only expert knowledge; the second (2.2) requires expert knowledge and data; the third (2.3) requires expert knowledge, data and models.

►► 4.2 Proposed solutions

4.2.1 Integrating qualitative expert knowledge for tree-based decision support systems

Cheese industries benefit from geographical indications and designations of origin, which supports their product differentiation strategy based on the promotion of local products linked to their terroir and experienced know-how in both production and processing. Internal changes within these types of quality schemes, particularly in terms of operator renewal and training, significantly undermine the preservation and transmission of this know-how. The development of “digital” methods enabling the use of knowledge bases opens up new perspectives in terms of promoting collective experience within a sector.

Based on qualitative expert knowledge, a unique method for capitalizing and supporting the recommendation of technological actions was developed. The methodology enables the collection of formal and informal knowledge from production line operators and experts on the relevant process, as well as the collective validation of this knowledge. The proposed software encodes this knowledge in a knowledge representation language based on a generic ontology, also using concept maps, process diagrams, hypermedia sheets, web pages, and downloadable documents for effective structuring and dissemination. All this information is then used to provide a semi-automated recommendation support tool.

The DOCaMEx platform, which was created to support the development of software to capitalize on and mobilize cheesemaking expertise and experience in sectors that promote their local terroir, consists of two interoperable web applications: CAPEX (short for the capitalization of experience; Buche et al., 2019, 2023) and the Knowledge Book (Baudrit et al., 2025), as represented in Figure 4.2.

- CAPEX is a semantic web-based reasoning engine whose goal is to suggest levers to correct a defect or achieve quality in a product in a food chain.

- The Knowledge Book is a semantic web-based hypertext platform designed for stakeholders to organize, share, and capitalize on knowledge.

Technical interoperability between the two applications is ensured through the use of standard HTTPS communication protocols and unified authentication and authorization mechanisms via a central authentication service server. Semantic interoperability is ensured through the use of a common vocabulary between the two applications, allowing CAPEX-suggested levers to be linked to the Knowledge Book resource that explains them. The whole process involves three data processing steps:

- Data structuring
- CAPEX explanatory view
- CAPEX action view

The data structuring step (already presented in Chapter 2, Section 2.2.2) allows users to build decision trees that represent technical reasoning. To build these trees, the method consists in interviewing technicians and experts about elements of reasoning able to explain standard appearance or product quality criteria obtention. Data structuring, which embeds explanatory mechanisms, also represents possible

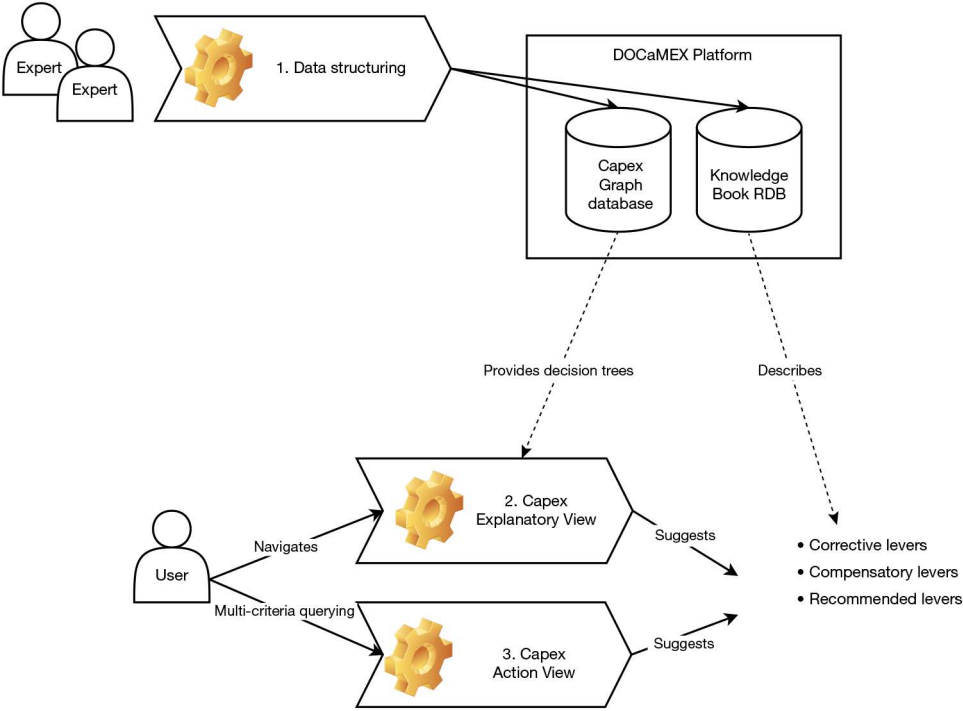


Figure 4.2. General workflow for the DOCaMEX platform. Cogs represent the data processing steps. Expert knowledge is collected and structured into two interoperable web applications, the CAPEX graph database and the Knowledge Book. The first provides decision trees that help users navigate through the information and action recommendations; the second describes the suggested levers based on the user’s navigation and querying.

cause-and-effect relationships between descriptors (defects and qualities) and intervention levers. These levers can be corrective, to correct a defect; compensatory, to reduce the impact of a defect; or recommended, to achieve or maintain a quality.

Technical reasoning is represented by possible causal relationships that exist between product defects and/or characteristics and technological actions, based on explanatory mechanisms. The tool makes it possible to determine, using a multicriteria approach, the recommendations for technological actions that are best suited to a given situation, based on technological efficiency indicators derived from the collection and formalization of expert knowledge.

The decision support system consists of two complementary access modes to the knowledge base content, i.e., the explanatory view and the action view. Both views can be used independently or together, depending on the application.

The explanatory view displays the decision tree associated with a given situation of interest, including all possible explanatory situations, associated analytical parameter values, and technical actions to correct/achieve the situation of interest. Using the explanatory view, a systematic review of all possible explanatory situations is performed.

As shown in Figure 4.3, the descriptor “Excessive Salting” (characterized by the rate of NaCl in red) is linked to a tree of explanations and situations of interest.

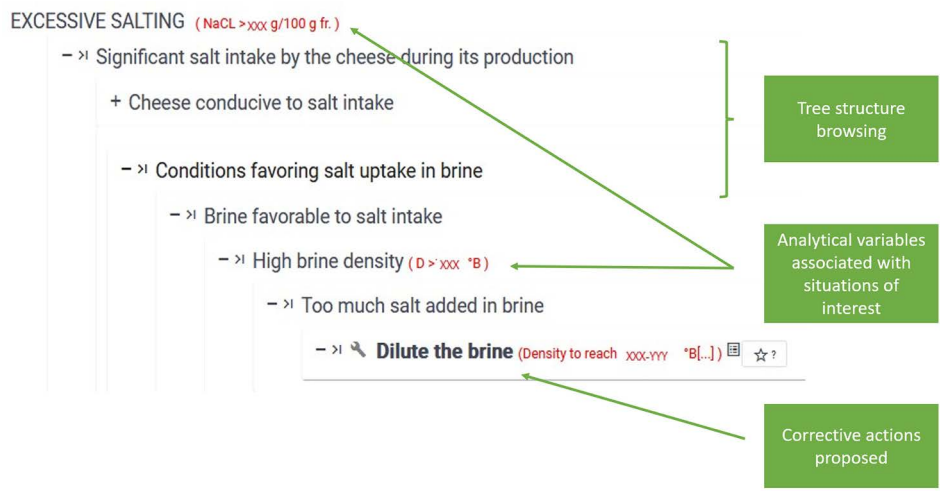


Figure 4.3. CAPEX screenshot: explanatory tree for the descriptor “Excessive Salting”.

Each item in the tree can be linked to a resource in the Knowledge Book, such as a causal map or a fact sheet, so that users can access scientific information via a hyperlink.

The action view displays the list of all actions/recommendations associated with a given descriptor. Figure 4.4. shows an example. It allows multicriteria filtering, ranking of actions, and identification of side effects. Solving a contextualized problem is done using the action view through the multicriteria filtering mode, sometimes complemented by the explanatory view.

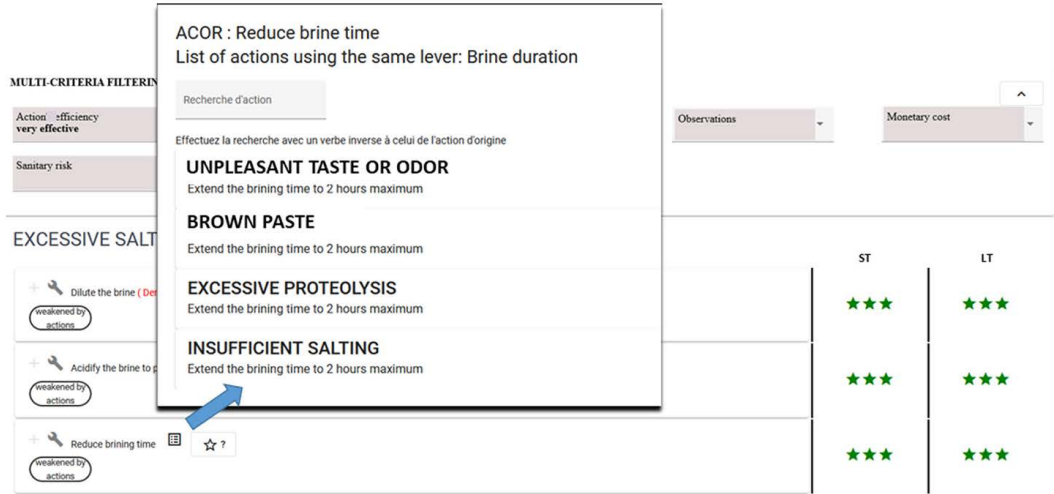


Figure 4.4. CAPEX screenshot: example of very effective actions filtered using the action view related to a given descriptor – in this case, the “Excessive Salting”. Clicking on the action “Reduce brining time” identifies a list of possible side effects (shown in the upper part of the window) that may occur if the “Brine duration” lever is used.

Multicriteria filtering allows users to quickly and easily select actions that meet their specific needs. It is intended for users who want to filter actions according to different contexts, i.e., a set of criteria that are of interest to them. The filtering process involves selecting one or more filters – such as cost, health risk, or technological efficiency – and applying them to a list of actions. The filtered list can then be sorted and paginated for easier navigation.

One of the main objectives of multicriteria filtering is to minimize the time spent searching for specific actions. This feature allows users to quickly filter out irrelevant actions, thereby increasing productivity and efficiency.

Filters are dynamically generated based on the data in the current decision tree. This means that the available filters change depending on the context in which the user is working. In addition to filtering by individual criteria, users can combine multiple filters using logical operators. This makes it easy to perform complex filtering scenarios.

The filtered list can be displayed in several ways, including a grid view or a tree view. The grid view displays all filtered actions in a table format, while the tree view displays them in a hierarchical structure. The hierarchical structure displays the recommended action in a contextual decision tree.

The tool also provides information on the potential impact of a recommended technological action on other defects and qualities, and can be used for training.

4.2.2 Machine learning approaches for process modeling using databases structured by ontologies

Thanks to their adaptivity, machine learning approaches are well suited to process modeling. In this chapter, we consider a given transformation process where data has been acquired (for instance, manually from experimentation or automatically through sensors) and used to train machine learning algorithms. The trained model can then be used for prediction and generating new knowledge for a better understanding of complex processes that are hard to model. For instance, Perrignon et al. (2025) used a dataset collected on-site at an industrial cheese plant to train a model for predicting cheese dry matter, giving control parameters that cheesemakers can influence to optimize their output.

In this section, we present the process of training machine learning models guided by qualitative expert knowledge. This approach is structured around three main steps, illustrated in Figure 4.5:

- Knowledge structuring and reliability assessment
- Deductive and abductive reasoning
- Capitalization of the models

The interest of this approach is that while these three actions are all tied to the training of machine learning models, they are also independent and can be updated at any moment during the process.

Knowledge structuring consists in gathering all potentially relevant data, be it from external sources (open-access databases, scientific articles, etc.) or internal (experimental results, etc.), and pooling them under a common structure based on a

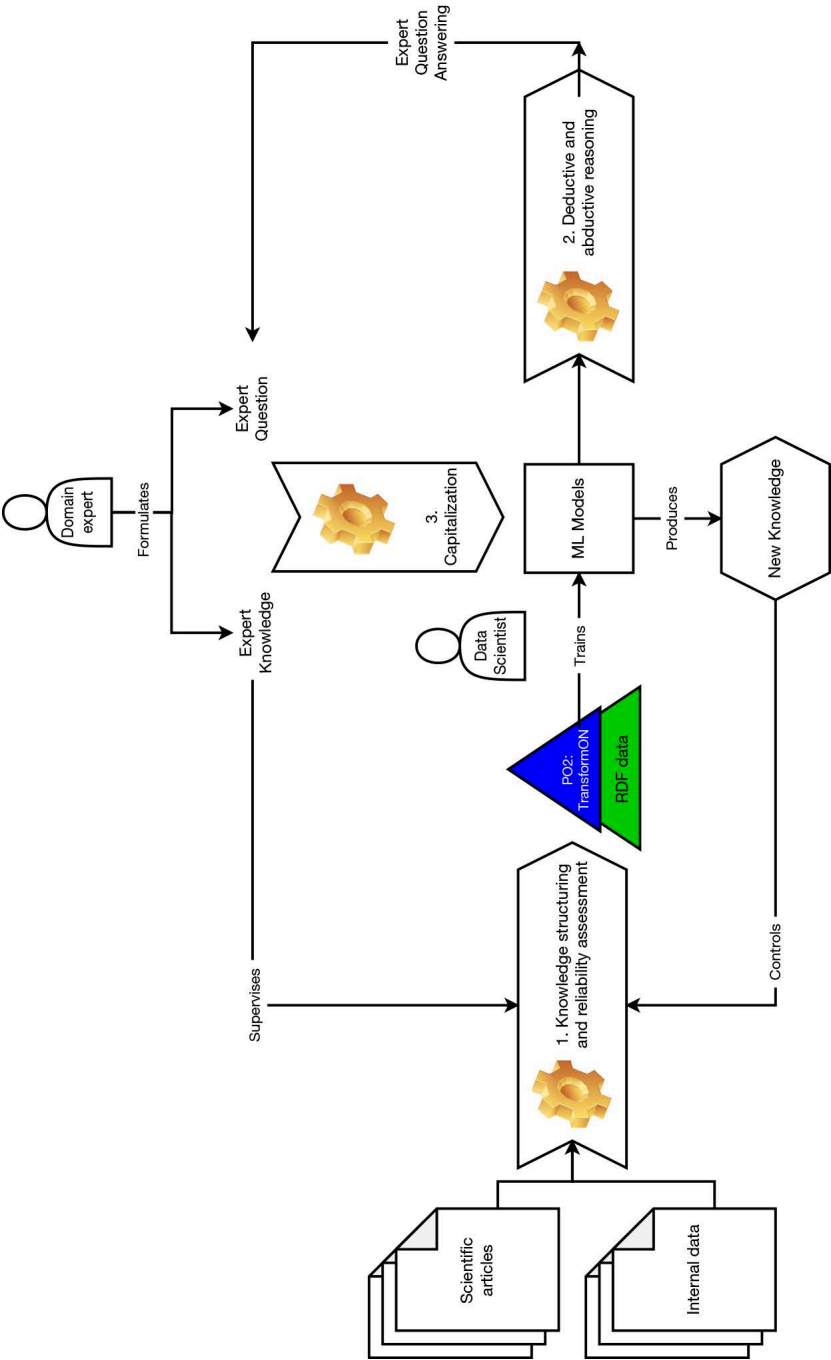


Figure 4.5. General workflow for the use of machine learning (ML) approaches for process modeling. Cogs represent the data processing steps. The blue triangle represents an ontology. The green triangle represents the knowledge base. Expert knowledge is first structured into an ontology, then used to train machine learning models to feed the deductive and abductive reasoning process. The results can then be used to answer expert queries and support knowledge capitalization, supervised by the data scientist.

controlled vocabulary. In that section, the TransformON (Weber et al., 2023) ontology is used: readers who want to read more about possible structural models (including TransformON) are invited to go to Chapter 2.

The TransformON ontology adapts well in this case: given the same goal (the transformation process), it allows the user to browse the different repetitions under different conditions (the itineraries). Since machine learning relies on repeated measurements to identify statistical relations, this particular structuring provides a good overview of the possible existing variables, their combinations, and whether their quantity is high enough or not.

This last part falls under data reliability assessment: estimating reliability is crucial for scientists who want to reuse data to draw new conclusions. Research question Q2.4 discusses this topic in Chapter 2 and proposes a method for assessing data reliability based on meta-information. Multiple criteria are used, each contributing a different perspective on the data's reliability. In a broader sense, this step can also be controlled manually by expert knowledge to check if the domain is well described, if the values are representative, or if there are not too many missing values or variables. Particular attention must be paid to spurious correlations: the experimentation plan must be equilibrated, and all variables must be controlled to prevent the model from learning the wrong statistical relations.

Once the data has been structured and validated, models can be taught by training them on machine learning algorithms, i.e., fine-tuning their parameters to best represent the data and allow prediction tasks.

This task can be automated using the established knowledge structuring. SPARQL queries, for instance, can be used to interrogate an ontology and automatically generate a training dataset. This makes it easy to adapt to any database update: in the context of the transformation process, where data is typically produced continuously, this allows machine learning models to be kept up-to-date: every time a new itinerary is added to the ontology, the model can be retrained.

Deductive and abductive reasoning consists in addressing specific expert questions.

Experts can present two kinds of questions about the domain: competency questions (CQs, which aim to reason about the domain, covered in Chapter 2), and causal knowledge questions (cKQs, which aim to reason about causal dependencies), illustrated in Figure 4.6. Once a model is set, reasoning can be used to answer the cKQs. Given a generic machine learning model f , two kinds of cKQs can be distinguished:

- cKQ1: Given a set of values for X and a trained model f , we predict the most likely value y for $f(X)$. Answering cKQ1 requires deductive reasoning. This approach was chosen, for instance, in Munch et al. (2023b). In the context of processes, these questions can, for example, aim to predict the outcome of specific control parameters.

- cKQ2: Given an objective value for y and a trained model f , we predict the most likely set of values X so that $f(X)=y$. Answering cKQ2 requires abductive reasoning. This is the approach chosen, for instance, in Munch et al. (2022a). In the context of processes, these questions are used for reverse-engineering purposes: given a desired outcome, users would like to find the required control parameters to achieve it.

The principal difference of the two approaches lies in the diversity of possible answers: deductive reasoning has one output y (usually with an associated confidence). Thus, deductive reasoning is a forward approach: the model makes a prediction.

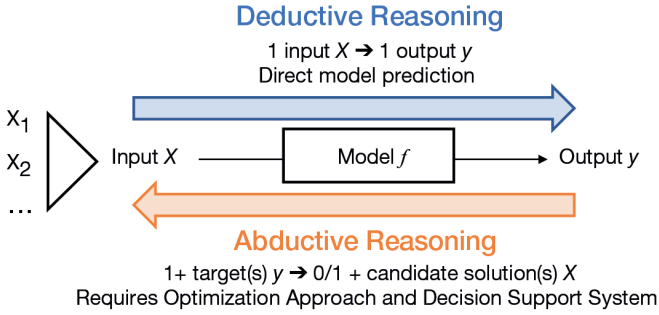


Figure 4.6. The two kinds of causal knowledge questions covered in this chapter. While deductive reasoning produces a single answer (y), abductive often generates multiple answers due to X being a set of values ($x_1, x_2, \dots$). Given the possible combinations, several X might be possible to obtain the desired output y .

By contrast, abductive reasoning can result in multiple possible combinations for the different variables in X . It is a backward approach: since reaching the target y is not always possible, it generates a set of solutions that are as close as possible to it: there can be one answer, multiple answers, or no answer. To do this, optimization methods can help search for the most appropriate set. Depending on the model's complexity, the proposed set of solutions can be incomplete, meaning not representative of all the possible X that can lead to the target y . In addition, optimizing becomes a more complex problem when multiple models are combined, with potentially conflicting targets: if y_1 decreases when y_2 increases, how do you generate a set of X that maximizes the values of both y_1 and y_2 ? To answer this question, Belna et al. (2022) proposed the use of multi-objective optimization to generate novel operating procedures for milk microfiltration that maximize yield while decreasing cost. An overview of such methods is provided in Konak et al. (2006).

Abductive reasoning usually generates a pool of potential candidate solutions that need to be refined. To do so, a multicriteria decision support system can help select the optimal solution(s). Perrignon et al. (2026) proposed, for instance, using the TOPSIS method (Hwang and Yoon, 1981) applied to potato chip production. Using three different trained machine models for different quality variables of potato chips, TOPSIS allows the user to rank solutions by preference criteria selected by experts (fat content, moisture content, and lightness).

Capitalization entails analyzing the result of deductive and abductive reasoning. Two cases are possible:

- The cKQs cover what is already described in the domain (e.g., existing partial models, datasets). In this case, the expert can validate the model or not and suggest ways to improve it (e.g., add new data, weight some points).
- The cKQs do not cover established knowledge. In this case, their results can be used to generate new knowledge. This new information can be used to better understand the domain, suggest new experiments to verify the model's prediction, or better control the quality of future data integrated into the knowledge base, among other things.

Both approaches to reasoning are complementary: deductive reasoning can be used to verify whether a model is trustworthy, while abductive reasoning allows users to discover new information about the domain. Both can also benefit from explainable AI methods.

With the advent of machine learning approaches, the need to explain predictions has given rise to the field of explainable artificial intelligence (XAI). In order to avoid “black boxes” whose outcomes are difficult to interpret (Vilone and Longo, 2021), different methods have been developed to better understand the decisions made by a trained model. An overview of such methods is presented in Ali et al. (2023). Munch et al. (2023b) applied the SHapley Additive exPlanation (SHAP) (Lundberg and Lee, 2017), an explainable method based on game theory, to answer cKQ1: Given the composition of a food product and its storage temperature, what is the CO₂ solubility? The method is model-independent and calculates the importance of each predictive parameter for the prediction value, giving the user insight into the importance of the parameters. In the case of CO₂ solubility, the method highlighted the influence of temperature on the composition’s importance during prediction.

4.2.3 Fuzzy set querying to combine multiple expert deterministic models

The solution proposed in this section is based on three different actions, described in in Figure 4.7: (i) extraction and transformation of the raw data into a structured database, (ii) simulation based on an existing deterministic model to optimize the database research, and (iii) bipolar multicriteria flexible querying.

The first two actions remain very similar to the one described in Section 4.2.2.

Extraction and transformation are part of a data gathering and structuring step used to bring all relevant data together under the same vocabulary. This step is crucial to address the multiple objects and issue of multiple scales often encountered when dealing with transformation processes. In particular, it usually allows for data from different sources to be combined for capitalization. Interested readers can refer to Chapter 2 for a full description of both tools available for this approach, the TransformON ontology and the naRyQ ontology. In this section, this step creates a database that can be automatically queried by the solution.

Simulation relies on the availability of one or more deterministic models – that is, models capable of producing a same output given specific input parameters. These models may take the form of physical equations or data-driven approaches (the training of such models is discussed in Section 4.2.2). They form the foundation of the approach presented in this section: given access to a deterministic model suitable for simulation, how can an expert efficiently search a database of candidate solutions to identify the one that best satisfies their wishes and constraints?

The bipolar multicriteria flexible querying approach presented in this section requires three inputs: (i) information about potential candidates from a database, (ii) results from the deterministic model simulation, and (iii) preferences and constraints from the expert. The output is a ranked list of all the possible solutions that should fit the expert’s query from the most to the least relevant.

To rank the solutions, the proposed flexible querying system must be able to deal with both the expert’s queries and the potential uncertainty in the queried database. To express such imprecision, fuzzy sets are introduced. Defined by Zadeh (Zadeh, 1965), fuzzy sets express the gradual membership of an element to a set, by generalizing a crisp set in which membership grades (representing desirability for expert’s

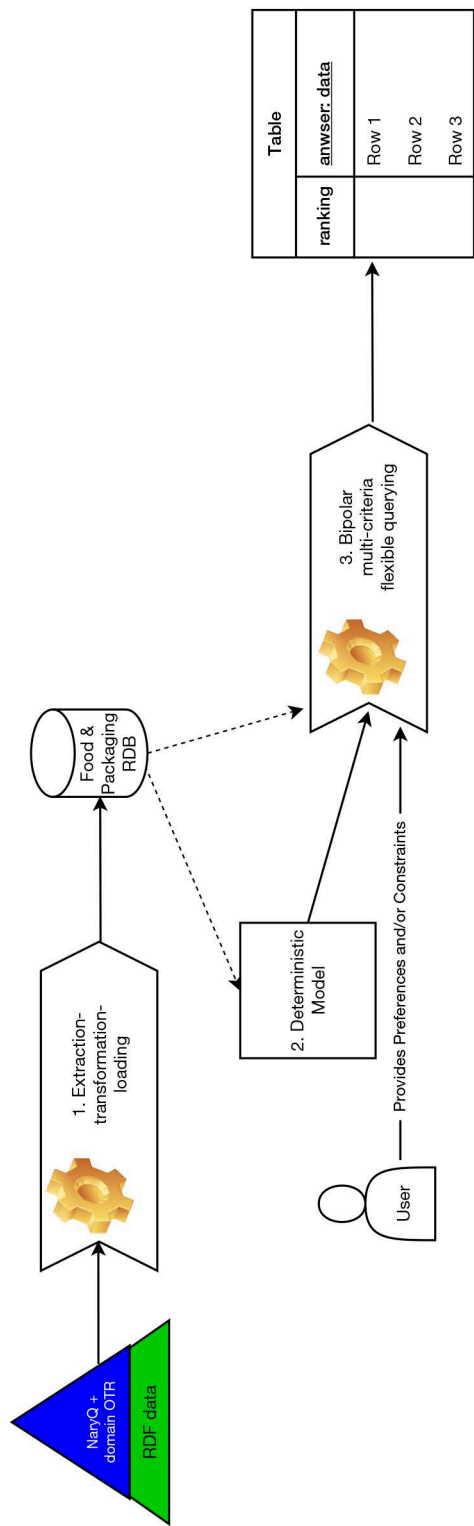


Figure 4.7. General workflow for fuzzy set querying with multiple expert deterministic models. Cogs represent the data processing steps. The blue triangle represents an ontology. The green triangle represents the knowledge base. First, data is extracted from the knowledge base to create a database. Second, the food parameters are used by the deterministic model to simulate the food characteristics that will inform the packaging selected. These results are then combined with the user's own preferences and constraints to suggest the most suitable packaging from among all the candidates in the database.

constraints or wishes, uncertainty for the information stored in the database) are between $\{0,1\}$. Figure 4.8 shows three examples of fuzzy sets defining preferences (desirability) as optimal values calculated by deterministic models for packaging: two constraints, O_2 and CO_2 permeances, and a wish, the optimal price. The combination of the three defines a bipolar query.

More formally, a query Q is defined as $Q = (C, W)$, where C is a set of different constraints $\{C(1), \dots, C(n)\}$ and W a set of different wishes $\{W(1), \dots, W(m)\}$. Elements are organized within sets such that constraints take priority over wishes. This allows the fuzzy set theory to be adapted to the expert's queries.

In the case of the database of solutions, the fuzzy sets are adapted to a semantic for imprecise data defined in Dubois and Prade (1997). For example, an unknown value is represented with a fuzzy set in which all elements belonging to the set are associated with the value 1. This way, both experts' preferences and the database's uncertainty share a common representation with fuzzy sets, allowing them to be combined using two classical measures of possibility theory (Dubois and Prade, 1988):

- The possibility degree of matching corresponds to the maximum degree to which the two fuzzy sets (expert's preferences and database uncertainty) overlap. It represents an optimistic evaluation of how well the database's candidate solution could satisfy the expert's preferences.

- The necessity degree of matching corresponds to the minimum level to which all elements of the database's fuzzy set are included in the expert's preference set. It reflects a pessimistic evaluation of how strongly we can be certain that the database's candidate solution conforms to the expert's preferences.

The higher both degrees are, the more confident we can be in the match quality. Based on these two measures, a ranking of the different solutions is provided: the top-ranked ones best satisfy the expert's preferences under uncertainty, while the lower-ranked ones still meet the constraints but with less confidence in the match.

The ranking based on both degrees allows for the handling of missing data: when information about a candidate solution is missing, the system assumes by default that the candidate could potentially satisfy all the expert's requirements – resulting in a possibility degree of 1. However, due to the lack of evidence, the system also cannot confirm this match, leading to a necessity degree of 0. In other words, this candidate solution verifies an open-world assumption: the absence of information is not evidence of failure.

Finally, in the case where every single candidate has a degree of matching of 0 for at least one constraint, the system warns the user and switches to candidates that are the closest, according to some distance, to satisfy the constraints.

► 4.3 Outcomes

4.3.1 Integrating qualitative expert knowledge for tree-based decision support systems

Two protocols were implemented to evaluate the efficiency of the proposed solution: one for CAPEX, and one for the Knowledge Book.

The general CAPEX evaluation was assessed in Buche et al. (2023). In a technological reasoning task, when faced with a specific situation of interest (whether aiming

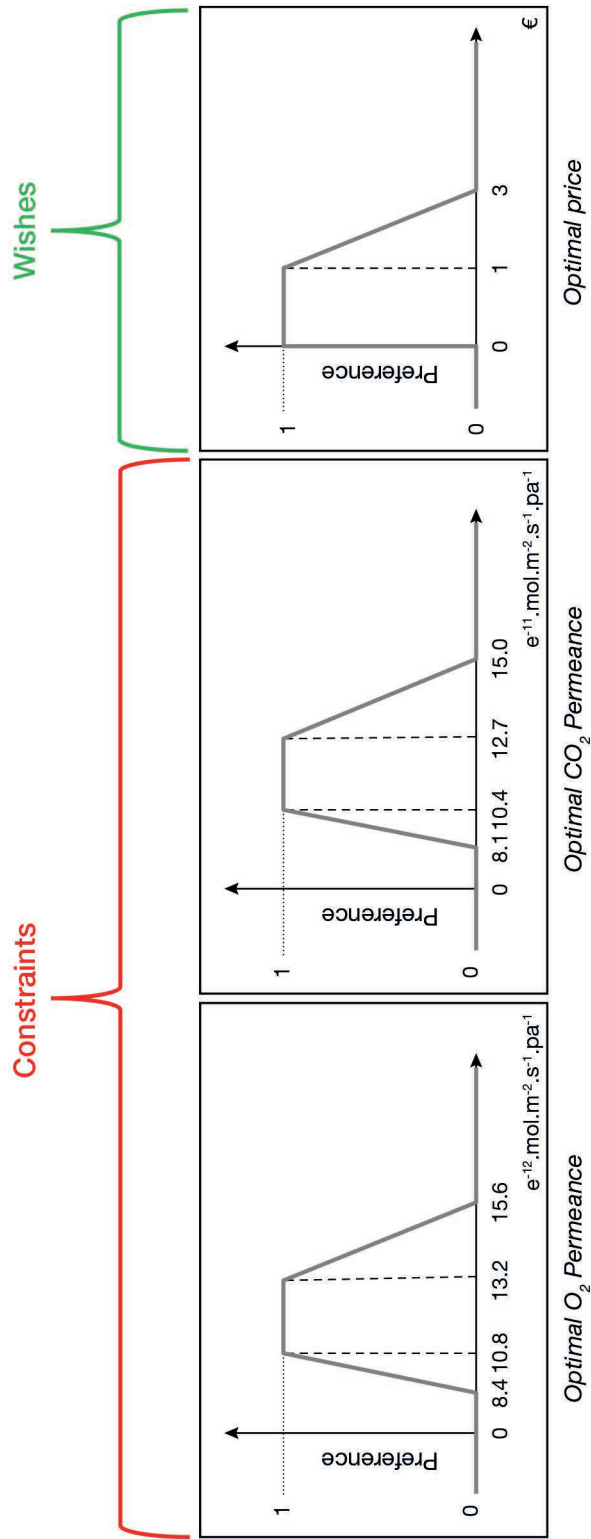


Figure 4.8. Example of a bipolar query using three fuzzy sets: two as constraints (for O₂ and CO₂ permeances), and one as a wish (for optimal price). The higher the preference, the more desirable the value is. For instance, the best prices would be between 0 and 1; prices between 1 (not included) and 3 becomes less and less desirable, and prices above 3 (not included) are unacceptable. Adapted from Guillard et al. (2015).

for a desired quality or addressing a defect), a technician must be able to consider all possible explanatory scenarios and the associated analytical parameters to verify whether this situation will occur. Additionally, the technician should know the recommended technological action to take. The protocol described by Buche et al. (2023) was designed to evaluate how the use of a multicriteria decision support system (MCDSS) influences performance in this reasoning task.

Fifteen participants, from three different cheesemaking production chains, took part in the test. Without the MCDSS, participants correctly identified 39% of the technological actions; with its use, this figure rose to 66.5%, representing a 27% improvement.

Only two chains (10 participants) completed the analytical value tests, as the first chain concerned generic knowledge typically taught in technical schools and not linked to specific analytical values. Since analytical values depend heavily on the particular cheesemaking process, it was not possible to assess this generic knowledge in that part of the test.

In the analytical value test, the 10 participants achieved 18.33% correct answers without the MCDSS, and 76.25% with it – a 60% improvement.

Both tests demonstrated a clear benefit – substantial in the first case, and very significant in the second – from using the MCDSS in supporting technological reasoning.

The electronic Knowledge Book (eK-Book) was assessed through its knowledge transfer capacity evaluation. Transfer is usually considered as the sum of transmission and absorption, i.e., its effectiveness is tied to both the support's ability to transmit the information and the user's ability to integrate the new notions.

To evaluate the knowledge transfer capacity of the cheesemaking eK-Book, three measures were computed in the form of a single-selection multiple-choice form:

- Cognitive load, which rates the mental effort users had to make to understand the concepts;
- Disorientation, which measures the ability to navigate and find the information needed;
- Understanding, which rates the gain in knowledge before and after navigating the eK-Book.

Lower cognitive load and disorientation indicate easier transfer, while higher comprehension indicates more effective transfer.

Cognitive load and disorientation are measured through a self-rating technique based on a seven-point mental effort rating scale (Paas et al., 2003), ranging from 1 (very low) to 7 (very high). Meanwhile, understanding is evaluated through the comparison between two tests, one taken before navigating the eK-Book (pretest), and another taken after (post-test).

The overall evaluation session is divided into different parts (Baudrit et al., 2024). Participants are first asked to complete an information sheet about their level of expertise, and then perform an initial test about the subject (pretest). Then, two scenarios are tested:

- Free navigation: Participants can freely navigate the eK-Book for 20 minutes.
- Scripted navigation: Participants must navigate the eK-Book to find requested information in order to illustrate the usability and efficiency of the eK-Book.

After each scenario participants perform a cognitive load and disorientation test. They must also pass a second test (post-test) after the scripted navigation.

Figure 4.9 presents the results obtained for the three tests for the 15 participants: correct answers before and after navigation (a), average disorientation and cognitive load (b), average difficulty of scenario-based navigation experienced before and after navigating the eK-Book (c). Participants came from diverse backgrounds (including operators, students, trainers, etc.) across three sectors (dairy school and two PDO cheese sectors).

Figure 4.9a shows, on average, a significant gain in knowledge: participants from all backgrounds show an increase in post-test scores compared to pretest scores, reflecting enhanced knowledge levels. Overall, the average score for knowledge acquisition increased from 11.1 out of 20 before using the eK-Book to 13.9 out of 20 after using the eK-Book, an improvement of 2.8 points. This indicates effective knowledge identification, understanding, and acquisition while navigating through the eK-Book.

Figure 4.9b shows that participants suffered minimal cognitive load and disorientation, with mean values of 2.2 and 1.95, respectively, corresponding to low to slightly low levels.

Similarly, Figure 4.9c shows a low level of difficulty in accessing requested information in the scripted scenario.

While these tests demonstrate the tool's effectiveness in supporting technological reasoning and its capacity for knowledge transfer, the deployment of DOCaMEx was shown to increase the skills of those involved in collecting and formalizing knowledge and know-how. All, regardless of their level of expertise, have reported an improvement in their skills in their field. Moreover, this task force of experts from different cheese production chains consolidated a large set of generic decision trees. This robust knowledge may be specialized and adapted to particular cheese product chains. As such, this knowledge engineering approach contributes to an overall enhancement of the knowledge representation in cheese production chains.

Another interesting aspect to note is the platform's use in educational settings. Today, the platform is used by national dairy schools in France and technical centers for initial and continuing training. Educational scenarios offering marked paths within the tool allow for its application by all types of operators.

Finally, in the case of DOCaMEx, the cheese industry adopted research-based tools as business tools that are now integrated into current practices.

4.3.2 Reverse engineering for biocomposite packaging design

The work presented in this section is extracted from Munch et al. (2022c). It shows a way of combining expert knowledge (in the form of ontologies) and machine learning (as probabilistic models) for the design of biobased product conversion processes. A second iteration of this work has been carried out on an expanded version of the database in Munch et al. (2022a).

In this context, lignocellulosic fillers (LFs) obtained through dry fractionation of unrecovered organic residues (agricultural, urban, forestry, and from agrifood industries) are mixed at a certain ratio with PHBV biopolymer matrices to constitute biocomposites.

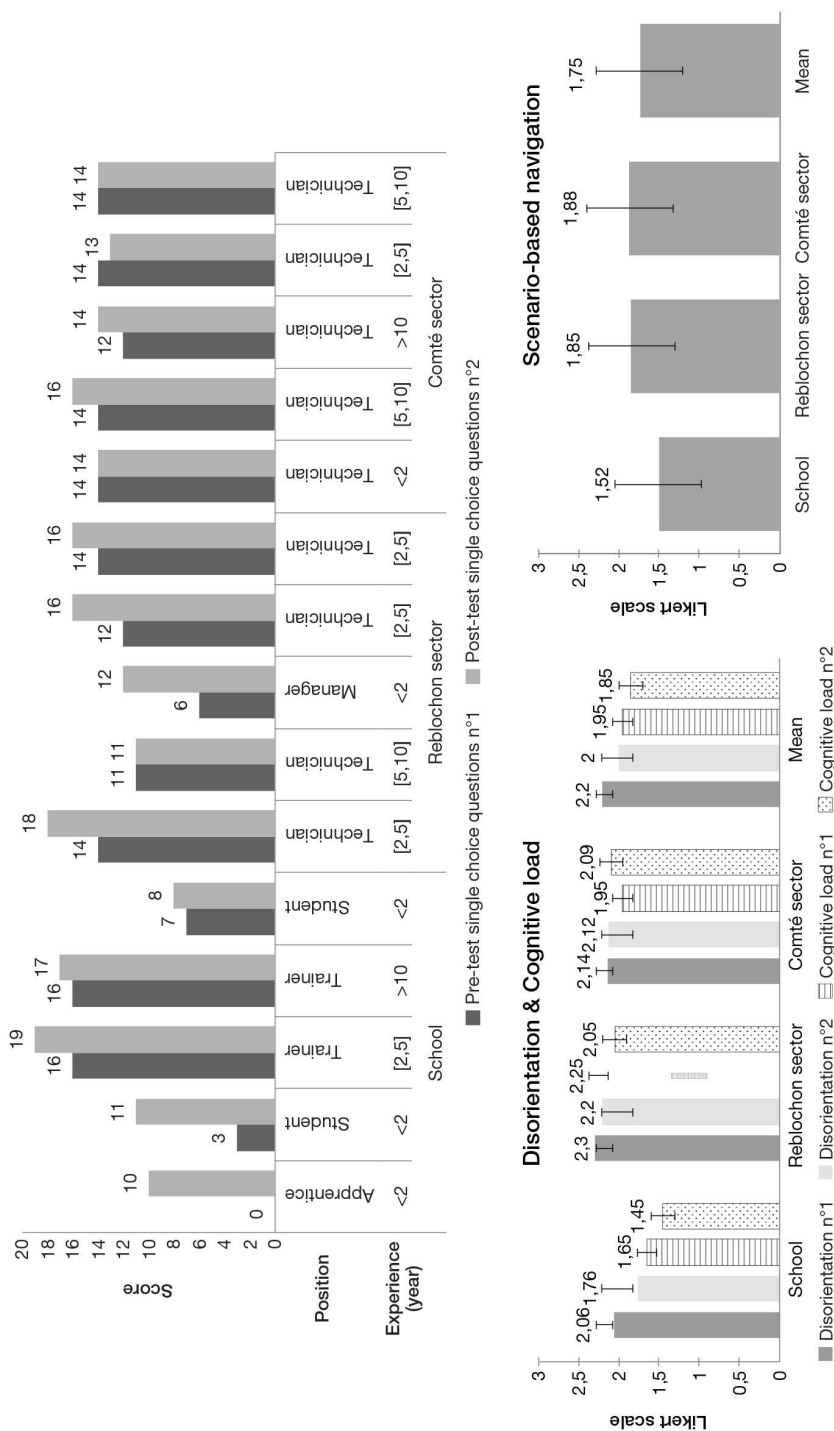
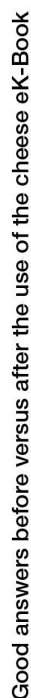


Figure 4.9. Test results of knowledge acquisition (a), mental effort (b), and the level of difficulty in retrieving precise information (c) in three cheese sectors (school and Reblochon de Savoie and Comté production chains). Figure from Baudrit et al. (2024).

The addition of LFs reduces the overall cost and environmental impact of PHBV-based composite materials (David et al., 2021), while modulating their functional properties (David et al., 2020). However, the introduction of lignocellulosic fibers has a negative impact over the biocomposite's brittleness and processability (Berthet et al., 2015): the higher the filler content, the more impacted they are. Therefore, the objective of this approach is to optimize the biocomposite's final functional properties while determining the most suitable LF and maximizing its quantity.

Following the approach described in Section 4.2.2, this application is carried out in four steps.

Knowledge structuring was performed using the TransformON ontology to structure datasets obtained in three different research projects (Munch et al., 2022b), presenting 70 original formulations with different LFs. Expert knowledge was mobilized to identify the key parameters to characterize (i) the LFs and (ii) the functional properties of the biocomposite. Particular attention was paid to the causality of the relations between the parameters: while there was no known model to express the relationship between the LFs and their filler content with the functional properties, experts knew which parameters could have an impact. Figure 4.10 illustrates this approach and the different groups of variables involved. After the relevant variables were selected, the resulting dataset included 84 itineraries, described by 15 different variables. It should be noted that feature engineering guided by expert knowledge was necessary here: since the 84 itineraries presented different PHBV bases, the final biocomposites' characteristics could not be compared directly. Instead, they were expressed as a ratio of the characteristics of the original matrix. Values below 1 indicate a decrease in the characteristic; the lower the value, the more negatively impacted the characteristic is.

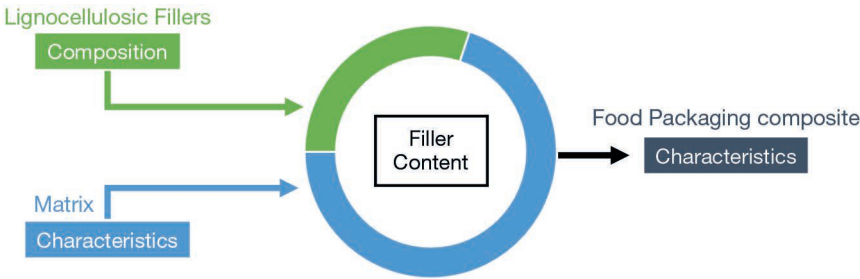


Figure 4.10. Categories of variables and their interaction in the model. Compositional parameters of the lignocellulosic fillers and filler content are used as model inputs to evaluate the deviation of the food packaging composite's characteristics from those of the matrix. Adapted from Munch et al. (2022a).

Once completed, a model could be trained (in this case, a Bayesian network). This approach is interesting because of its graphical component, which allows for easy-to-read results and facilitates the integration of expert knowledge. Bayesian networks are defined by two components:

- A graph structure, which describes the different variables (the nodes) and their relations (the edges). This representation is qualitative: it shows which variables have an impact over the others. Figure 4.11 shows the model trained in this example.

– Probability tables, which describe the interactions existing between the different nodes. This representation is quantitative: it indicates the probability of observing an event. Table 4.1 shows an example of the probability between two variables, filler content and peak temperature.

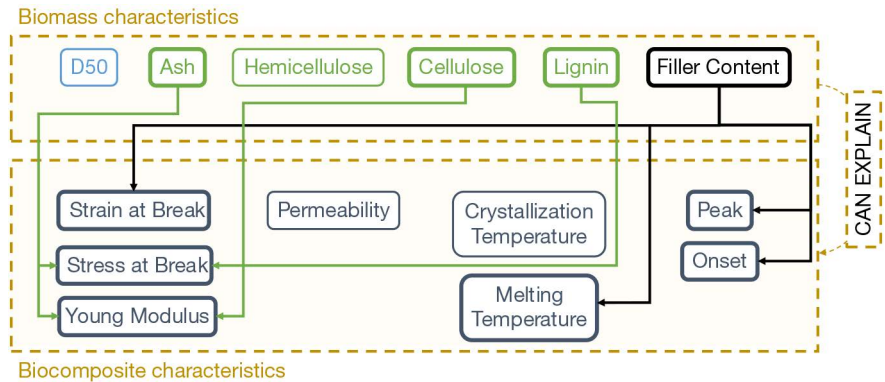


Figure 4.11. Learned Bayesian network. Variables are represented by rectangles (nodes), and relations by the arrows (edges). Expert knowledge constraints are indicated with the dashed lines: the six variables at the top can have a causal effect on the eight variables below, but not vice versa. Adapted from Munch et al. (2022c).

Table 4.1 Example of a probability conditional table between the two variables filler content and peak temperature. Bold probabilities indicate the maximum likelihood: for instance, if filler content is within]2;4], then peak temperature has a probability of 0.82 to be within]0.95;1].

Filler content	Peak temperature		
	]0.72; 0.95]	]0.95; 1]	]1; 1.16]
]2; 4]	0.09	0.82	0.09
]4; 11]	0.38	0.52	0.1
]11; 21]	0.91	0.09	0
]21; 50]	0.6	0.2	0.2

Reasoning was carried out in two steps. First, simple inductive causal knowledge questions were asked to verify the concordance with the state of the art, such as the influence of the filler content over the different variables. Table 4.1 shows that, in this case, an increase in filler content has a negative impact on peak temperature: the higher the filler content, the lower the peak temperature.

Second, inductive reasoning was used to generate possible lignocellulosic compositions that would best preserve the mechanical properties (in particular, stress at break and young modulus, since these characteristics only seem to be impacted by composition). By querying a biomass database, the model evaluated that both pine bark and rice husk had high probabilities (0.99 and 0.82, respectively) of preserving their mechanical properties.

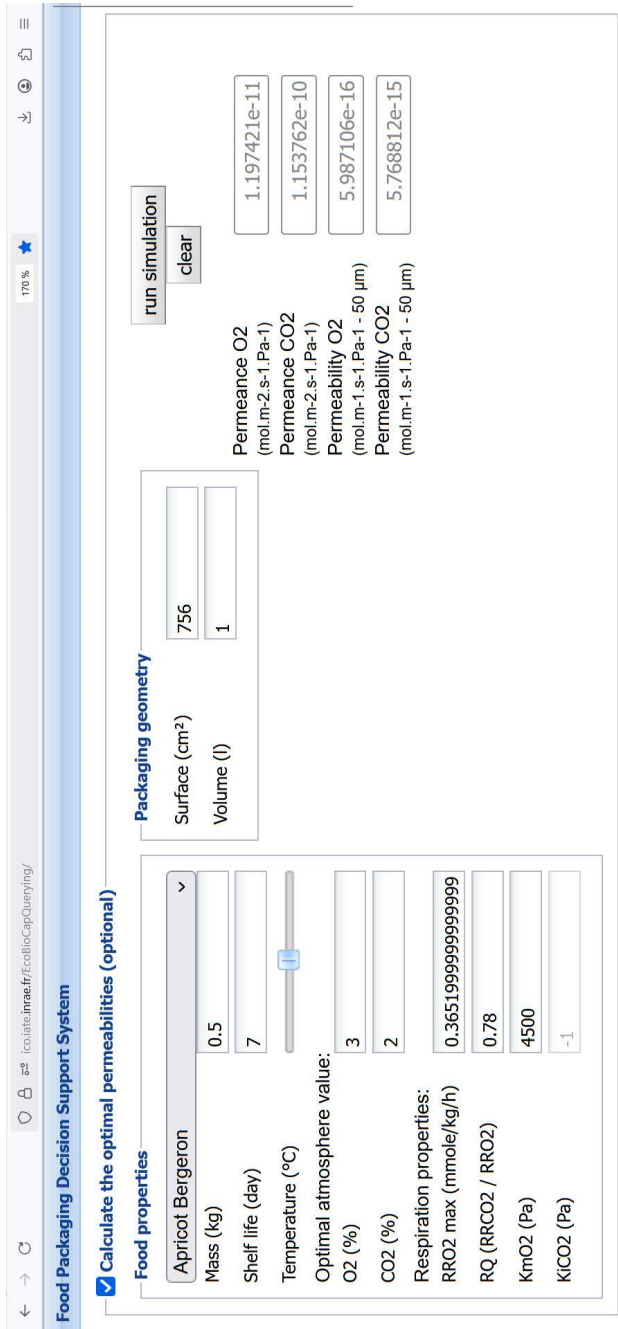


Figure 4.12. Virtual MAP module screenshot of the permeance simulation module. In this example, the permeance values have been calculated for “Apricot Bergeron.”

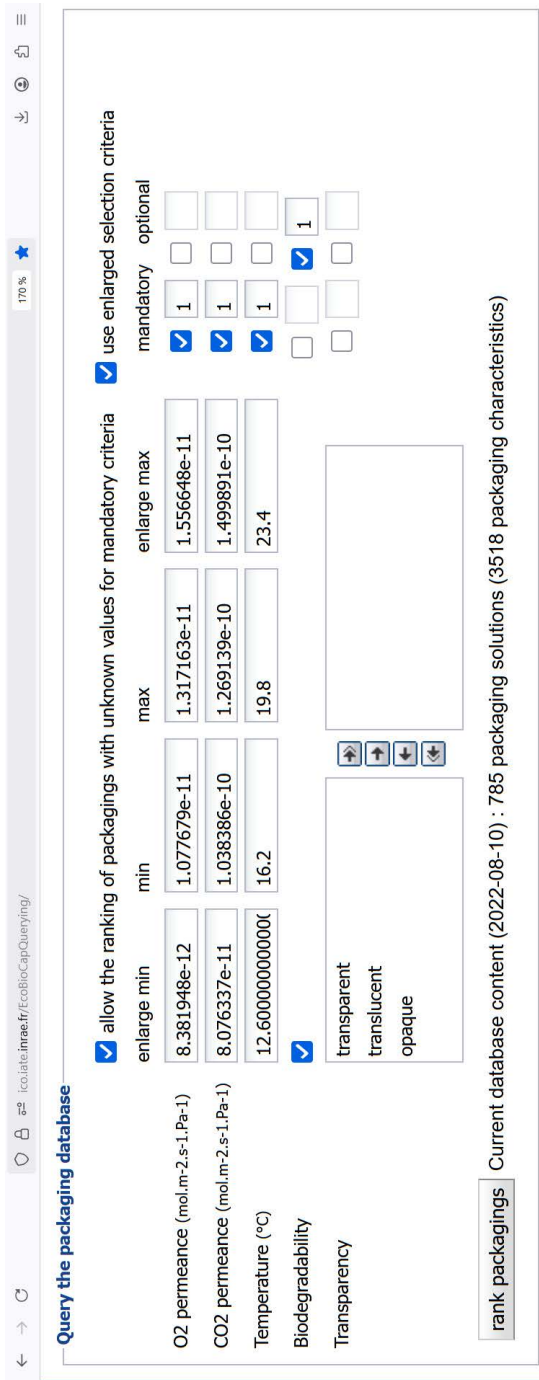


Figure 4.13. Virtual MAP module screenshot of the expert's indicated preferences. Some of the preferences (such as the permeances) are direct outputs from the simulation module. In this case, the expert has indicated three constraints (on permeances and temperature) and one wish (on biodegradability).

4.3.3 Fuzzy set querying to combine multiple expert deterministic models

Guillard et al. (2015) describe in this context a decision support system designed to recommend modified atmosphere packaging (MAP) for fresh products. The deterministic model used is a virtual MAP module, which simulates the packaged food's shelf life; the expert knowledge is represented by stakeholders' constraints and wishes for the final packaging. This system is also supported by two databases that gather information on both (i) potential fresh products and (ii) packaging material.

The main goal of this approach is to be able to deal with the ranking of expert queries based on their preferences and the sometimes imprecise values displayed in the database.

The interface of the implemented version is divided into three modules: (i) the permeance simulation, (ii) the expert's indicated preferences, and (iii) the ranked solutions.

Figure 4.12 presents the first module. It is connected to the fresh food database to retrieve the characteristics associated with the selected fresh food. In this example, it shows the permeances computed for the Bergeron apricot, for a shelf life of 7 days, in ambient temperature (18°C), with a food mass of 0.5 kg, a volume of 1 L, and a surface of 756 cm².

Figure 4.13 presents the second module. In this version, the user can only specify their preference for the CO₂/O₂ permeances, the storage temperature, the biodegradability, and the transparency of the packaging material. Permeances computed in the first module are autopopulated in this section, with a fuzzification on the extreme values to enhance the tolerance. The user can change these values later by hand. Wishes can also be expressed: in this case, the user wished that the suggested packaging might be biodegradable.

Figure 4.14 presents the ranked solutions based on the packaging database. In this example, the wish could not be fulfilled, and the highest percentage of known values is 75% for low-density polyethylene. This module can also display further details stored in the database (such as commercial or other packaging, or available references in scientific publications).

► 4.4 Conclusion and resources

In this chapter, three research questions have been addressed to reason about the integration of experimental data and technological expert how-know for decision support in process modeling:

- Q1. How can technological recommendations to solve a production defect or reach a given quality be determined using collective know-how expressed as qualitative knowledge?
- Q2. How can the absence of a deterministic model be managed from a reverse-engineering optimization perspective?
- Q3. How can we extract information from an incomplete data source containing candidate solutions using multicriteria queries while taking into account constraints and wishes for decision-making purposes?

Packagings ranking			
ranking	name	type	% known value
1	Low Density Polyethylene	Low Density Polyethylene	75
Id.: 620 Name: Low Density Polyethylene Type: Low Density Polyethylene Reference: Modified atmospher packaging of fruits and vegetables Authors: O2 permeance (mol.m-2.s-1.Pa-1): [9.437999999999999E-12, 3.146E-11] 🔗 (permeance computed with an estimated thickness of 50 µm) CO2 permeance (mol.m-2.s-1.Pa-1): [1.8634E-11, 1.863399999999997E-10] 🔗 (permeance computed with an estimated thickness of 50 µm) Temperature (°C): not available Biodegradable: false Transparency:			
2	LDPE/LLDPE/Organoday 3%	Polyolefin	50
3	Galactomannan C. pulcherima (1.5%w/v)/Glycerol (2.0%v/v)	Galactomannan C. pulcherima (1.5%w/v)/Glycerol (2.0%v/v)	100
4	(0.47%) Magnetic iron oxide / Zein	(0.47%) Magnetic iron oxide / Zein	50
5	Micro-perforated Plastic	Packaging	75

Figure 4.14. Virtual MAP module screenshot of the ranked list of suitable packaging, based on the results computed for the apricot and the expert's preferences. Details are given for the first proposed solution.

These three questions assume different levels of available knowledge, from qualitative (expert knowledge) to quantitative (existing models). Combining them makes it possible to consider the available data in different ways and to deal with possible imprecision and uncertainty.

Three solutions have been proposed to answer these questions and compared to the current state of the art. The software workflow implementation for each solution has been presented and illustrated. Available open resources (data, software, video, courses) have been listed. Main outcomes obtained using those solutions have also been presented. All of them have high technology readiness levels as they are all being used in production.

Future avenues for this work will be to develop interoperability between these solutions in order to design digital twins. In fact, while AI solutions are experiencing unprecedented development today, fueled by the promise of generative AI, this chapter has highlighted the major importance of expert knowledge integration in all modeling processes. Although most of the current algorithms allow a good grasp of the underlying relations between the different variables, Pearl (2009) highlighted the necessity of causal reasoning to address complex modeling issues. To propose actions or suggest corrections to a process, one must be able to reason at different levels: associations (“What is happening?”), interventions (“What will happen if I take action?”), and counterfactuals (“What would have happened if I had acted differently?”).

As such, being able to combine experimental data and technological expert know-how is crucial for decision support.

Available resources		Available resources			
Research question	Proposed solution	Main scientific papers	Data papers and datasets	Software codes	Videos
Q1	DOCaMEx CAPEX	Multicriteria decision support system based on decision trees (Buche et al., 2023)	Example of decision tree and associated efficiency indicators (https://doi.org/10.57745/SEJP1B)	CAPEX resources https://hal.science/hal-04957075v1	Courses and technical reports
Q2	POND	Combination of ontologies and machine learning approach (Munch et al., 2022c) Reverse-engineering approach (Munch et al., 2022a)	Dataset of biocomposites (Munch et al., 2022b)	https://hal.inrae.fr/hal-05310941v1	
Q3		Multicriteria decision support system based on a list of candidate solutions for food packaging (Guillard et al., 2015) Database querying when user preferences are bipolar (Destercke et al., 2011)		https://hal.inrae.fr/view/index/docid/5296718	Simple use https://hal.science/hal-05123333v1 Using qualitative & quantitative criteria https://hal.science/hal-05176679v1

Methods for Collective Decision Support that Consider both Preferences and Contradictory Arguments from Agrifood Chain Stakeholders

Pierre Bisquert, Rallou Thomopoulos

►► 5.1 Introduction

This chapter aims to explore argumentation-based decision support in the context of food chains. Specifically, we present the context and describe two complementary methods, DAMN and MyChoice, and their associated use cases related to biowaste by-product valorization. By focusing on the outcomes of these approaches for their respective use cases, we highlight their relevance and utility. Furthermore, the two methods are compared to examine their workflows, applications, and results. This comparison itself is a valuable result, offering insights into how such methods can be adapted to meet diverse decision-making needs in food chains.

New trends and developments in food chains typically require agreement and cooperation among various stakeholders, resulting in a need for collective decision-making methods. In this chapter, we explore argumentation-based decision-making, which relies on the provision of explanations to support decisions. This approach was developed to help decision-makers analyze, compare, and assess the acceptability of different alternatives (e.g., technologies, food processes, variants of a product), based on explanatory arguments drawn from various sources and stakeholders with respect to different criteria and aims.

Supporting decision-making in food chains is critical to fostering sustainable, efficient, and ethical practices. These decisions often involve diverse stakeholders with varied preferences, priorities, and justifications. Ensuring alignment and adherence among stakeholders requires their perspectives to be adequately represented. This representation is vital for ensuring the legitimacy and acceptance of decisions, which in turn enhances their implementation and impact. Argumentation-based approaches provide a robust framework for explicitly modeling and reasoning about these justifications, thereby improving both the fairness and transparency of the decision-making process.

Two illustrative use cases exemplify the application of argumentation-based decision support in food chains: (i) ArgCW-LCA (Sohn et al., 2020), based on the DAMN tool, and (ii) emergent biowaste conversion chain analysis (Vivas, 2022), based on

the MyChoice tool. In Sohn et al. (2020), DAMN is used to reason about arguments relating to how LCA criteria should be weighted, based on the preferences and justifications of experts. These criteria and weights are then used to support specific decisions about valorization routes, emphasizing a content-driven perspective on sustainability indicators. Conversely, MyChoice is used to directly evaluate the valorization routes themselves based on the metadata of pro/con arguments, such as aims, criteria, and authorship. Together, these examples demonstrate how argumentation frameworks can operate at different levels, from general criteria-setting to direct decision evaluation, showcasing their versatility and adaptability to diverse decision-making challenges in food chains.

This leads to the main research question:

– Q5: How can stakeholder input, both in the form of preferences and justifications, be integrated into complex, argumentation-based decision support systems?

This research question can be refined on more specific points:

– Q5.1: In what ways can justifications impact the evaluation of alternatives in collective decision-making in a real-world setting?

– Q5.2: How can argumentation-based tools shed light on the positioning of different stakeholders regarding possible decisions, based on various criteria?

While argumentation theory has produced a wealth of formal models, ranging from abstract frameworks such as Dung's argumentation framework (Dung, 1995) to more expressive structured (Besnard and Hunter, 2001) and bipolar (Cayrol and Lagasquie-Schiex, 2005) extensions, very few of these have been translated into usable software tools with the goal of supporting collective decision-making.

Research question Q5.1

To address real-world decision problems involving diverse and possibly conflicting stakeholder perspectives, tools must go beyond logical soundness and incorporate features such as:

- User-friendly interfaces that lower the barrier to participation for nonexpert users
- Multi-agent contribution support, allowing stakeholders with conflicting views to enter justifications in parallel
- Transparent and navigable justification structures, enabling traceability of decisions
- Semantic flexibility, allowing for different takes on a discussion

Most existing tools, such as ASPIC⁺ (Modgil and Prakken, 2014), Carneades (Gordon et al., 2007), DEFT (Hecham et al., 2017), DeLP (García and Simari, 2004), Flora-2 (Yang et al., 2003), AIPA (Delhomme, 2022), AFORI (Taillandier et al., 2023), and ArguMed (Saint-Dizier, 2012), were not built with collaborative or collective decision-making in mind, such as multi-actor and multicriteria analysis. These systems typically focus on individual or dialogical reasoning tasks, often hardcoding specific reasoning semantics such as ambiguity propagation or strict defeat. Moreover, many lack user-friendly interfaces and open-source access, and do not scale well in participatory settings.

To bridge this gap, we introduced Statement Graphs (Hecham et al., 2018), a flexible formalism that unifies various defeasible reasoning variants. Statement Graphs allow configurable semantics such as ambiguity blocking/propagating and team

defeat, enabling tailored reasoning strategies. This flexibility is especially useful in collective decision-making contexts, where different participants may rely on different reasoning intuitions.

Building on this foundation, DAMN (Bisquert et al., 2025) was developed not only to implement Statement Graphs but also to provide an interface that supports collaboration, justification traceability, and semantic configuration.

Research question Q5.2

Addressing this question requires the explicit representation, in the argumentation formalism, of the stakeholders expressing each argument, and of the criterion the argument relies on. Similar to the idea of “criteria,” some researchers have introduced values (Bench-Capon, 2003) or contexts (Bourguet et al., 2013) that extend Dung’s abstract framework, or viewpoints (Thomopoulos et al., 2015) in a logical framework. However, explicitly representing “stakeholders” is not typically addressed in formal argumentation systems. Going beyond the state-of-the-art on this aspect, DAMN collects information about arguments associated with different stakeholders, while MyChoice goes even further by providing multistakeholder analysis evaluating differential positionings of stakeholders regarding the alternatives, and their interests in the various criteria in their arguments.

►► 5.2 Proposed solutions

Argumentation-based decision support systems are designed to manage complexity and conflicts in decision-making scenarios by using formal argumentation frameworks. These frameworks facilitate the resolution of competing priorities and the justification of choices.

Different argumentation frameworks offer varying expressive power and structural complexity. Classical abstract argumentation (e.g., Dung’s framework) considers arguments as atomic entities in a graph of attacks. In contrast, structured argumentation explicitly models the internal logical structure of arguments, using premises and inference rules. Bipolar argumentation frameworks go further and introduce both attack and support relations between arguments, allowing a richer representation of reasoning dynamics. Both DAMN and MyChoice extend beyond classical models – DAMN through structured justifications and semantic labeling, and MyChoice by attaching metadata (aim, criterion, author, etc.) to arguments.

In some cases, argumentation-based methods feed directly into multicriteria decision analysis (MCDA) processes. For instance, research by Delhomme et al. (2022) or on ArgCW-LCA (Sohn et al., 2020), which will be presented later, uses argument assessment to derive the importance of criteria, which are then integrated as weighting into a TOPSIS ranking procedure (Hwang and Yoon, 1981). This highlights how argumentation can serve as both a deliberative layer and an input-generation mechanism for classical decision support models.

The core components of argumentation-based decision support include:

- Representation of arguments: identifying, collecting, and structuring arguments that support or oppose specific decisions
- Evaluation frameworks: applying systematic criteria to assess the validity and/or strength of arguments

– Decision workflow: guiding users through complex decision pathways by presenting clear justifications

These methods are particularly effective in food chains, where decisions must address multifaceted challenges encompassing environmental, economic, and social dimensions. By explicitly incorporating preferences and justifications, argumentation-based systems ensure inclusive and transparent decision processes.

5.2.1 DAMN

Presentation of DAMN

DAMN (which stands for Defeasible Argumentation for Multi-agent Negotiation) is part of a broader methodology for justified collective decision-making that integrates argumentation theory with computational social choice. At its core, DAMN uses the Statement Graph formalism, where each node (called a statement) represents a reasoning step, and directed edges denote argumentative relations (support or attack). The semantics governing these graphs allow nodes to be labeled as accepted, rejected, or ambiguous based on their interrelations and a rule-superiority structure. This flexible and defeasible logic-based foundation enables conflict-tolerant reasoning across diverse stakeholder justifications.

A simple but classic example is the so-called penguin problem (Figure 5.1), in which two agents hold conflicting beliefs about whether a penguin can fly. One agent invokes a general rule that all birds fly (a rule that may have exceptions), while the other asserts that penguins are birds that do not fly. Knowing that Tweety is an instance of penguin, two contradictory reasonings can be developed: (i) since Tweety is a penguin, it is a bird, thus it flies; (ii) since Tweety is a penguin, it does not fly. By expressing these

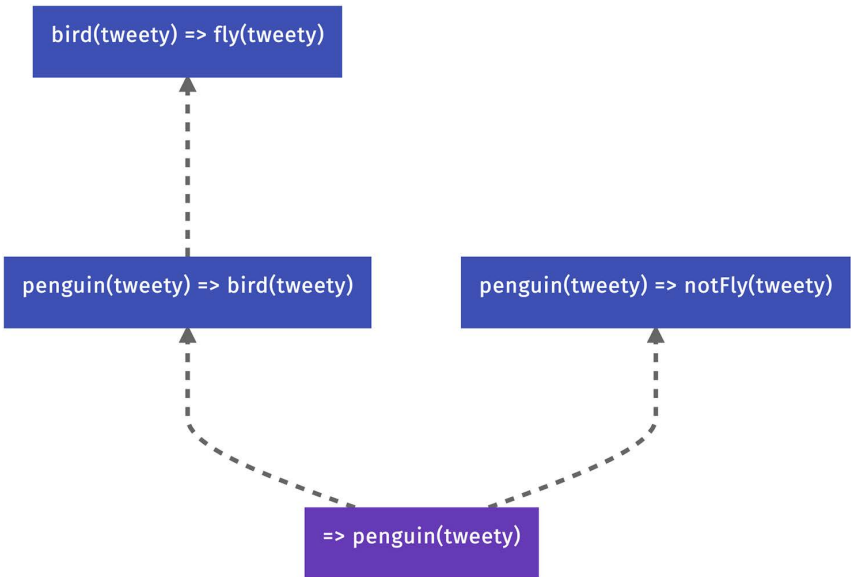


Figure 5.1. The penguin problem represented in DAMN. Dashed arrows represent support between statements, e.g., “Tweety is a penguin” supports “If Tweety is a penguin, then it is a bird.”

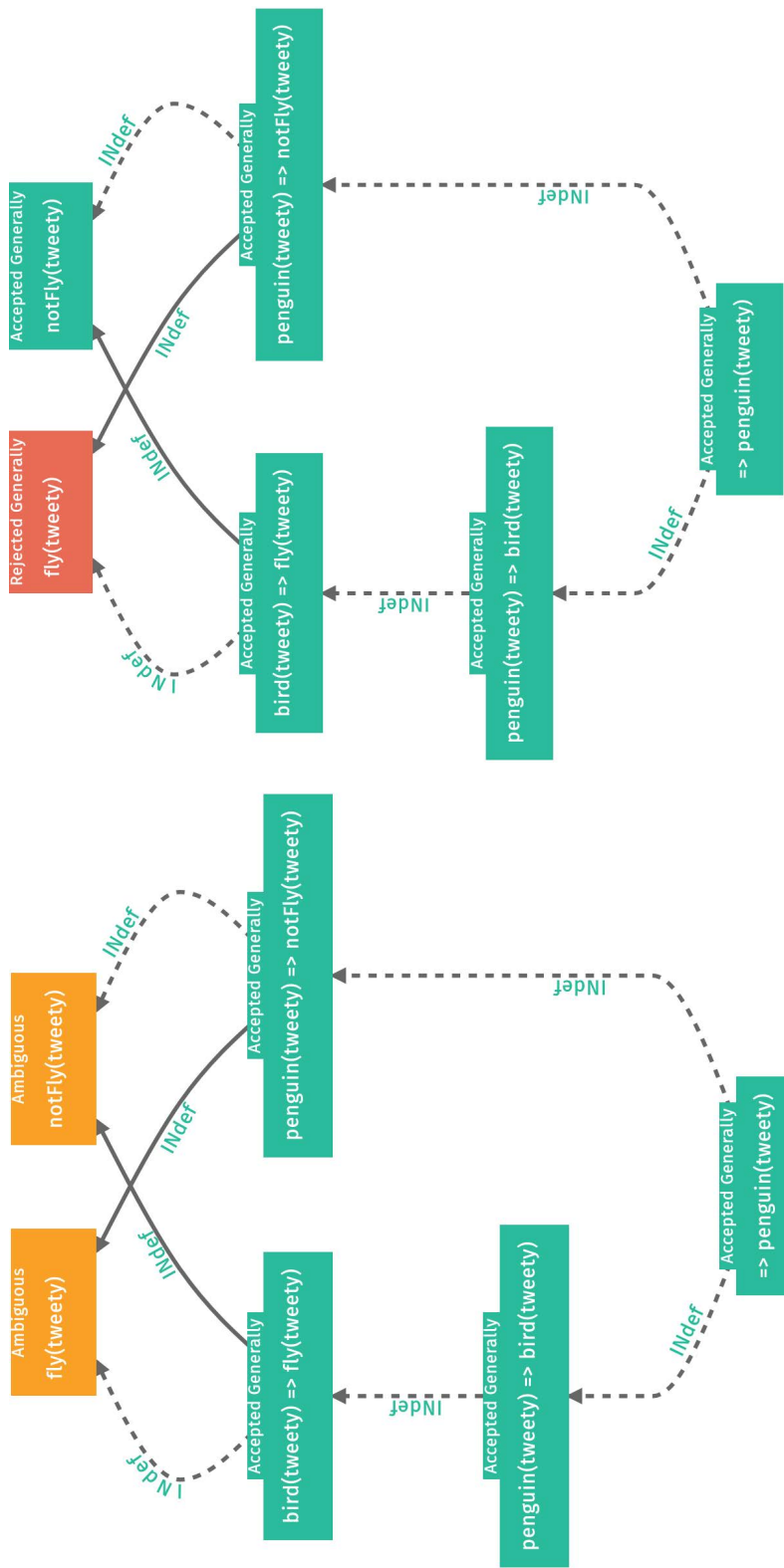


Figure 5.2. Illustration of the application of semantics to the penguin problem. On the left, applying any semantics leads to ambiguity about which statement should be accepted. On the right, by enforcing the specific rule that penguins do not fly, ambiguity is overcome and a conclusion can be reached.

arguments in logical terms,⁴¹ the system constructs a Statement Graph and computes logical interactions between arguments (attack or support):

One of DAMN’s notable strengths lies in its flexible treatment of ambiguity and defeat in argumentation reasoning (see more in Hecham et al., 2018):

- Ambiguity blocking versus ambiguity propagating: In ambiguity blocking semantics, when two arguments are in unresolved conflict (e.g., both are of equal “strength” or lacking superiority), ambiguity is contained, meaning ambiguous conclusions cannot be used to attack other arguments. In contrast, ambiguity propagating semantics allows ambiguous statements to influence further reasoning, spreading uncertainty through the graph through “ambiguous attacks.” DAMN supports both modes, letting users choose based on the nature of their decision context.

- Team defeat: DAMN also allows for “team defeat,” where multiple supporting arguments can collectively counter a set of attackers. This contrasts with systems requiring that each argument individually withstand all attacks. To a certain extent, team defeat reflects real-world collaborative defense of conclusions and enhances the expressive power of the reasoning process.

Applying a reasoning semantics allows us to compute argument status (accepted, rejected, or ambiguous). Depending on rule priorities, the tool concludes which justifications are accepted, rejected, or remain ambiguous, as shown in Figure 5.2. Indeed, without additional information, it is impossible to choose between both ambiguous conclusions (left graph). Adding the rule priority (penguins not flying is stronger than bird flying) allows the tool to provide a decision (right graph).

Figure 5.3 shows the typical four-step workflow that DAMN follows:

- Argument identification: collect justifications, arguments, and preferences of the stakeholders, typically through interviews
- Argument representation: formally represent collected arguments and preferences using an ontology relevant to the domain use case
- Content analysis: apply a reasoning algorithm to assess argument interactions (support and attack interactions)
- Decision modeling: assess argument status (accepted, rejected, or ambiguous) by applying defeasible semantics to inform or justify decision outcomes

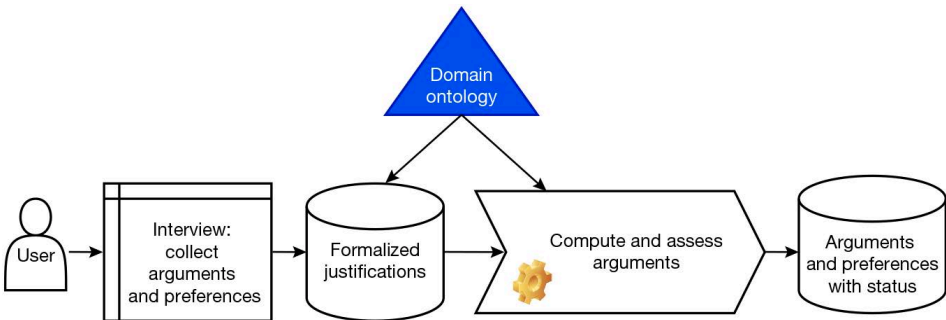


Figure 5.3. DAMN’s typical workflow, from users’ knowledge to formal representation.

41. While not absolutely necessary to understand the approach, interested readers will find a short primer on formal logic in Appendix 1.

Focus and key strengths: DAMN prioritizes the content and logical structure of arguments. It operates by explicitly modeling the relationships between claims and their justifications, allowing for detailed and structured evaluation of conflicting perspectives. This is especially valuable when decision-making must be grounded in substantive reasoning and the coherence of claims, such as in technical or sustainability-oriented contexts.

DAMN implementation

The DAMN tool is composed of modular components implemented primarily in Java (backend and reasoning module) and JavaScript (interactive web-based visualization). The tool is structured around two main modules (Figure 5.4):

- DAMN⁴² serves as the user-facing layer, allowing agents to input rules and facts. It supports browser-based interaction and visualization of Statement Graphs, making it accessible for nonexpert users.
- ELDR (Existential Logic for Defeasible Reasoning), a module of the Graal⁴³ logic-based reasoning framework developed by the Boreal team at the French National Institute for Research in Digital Science and Technology (INRIA), is the underlying reasoning engine. It applies defeasible reasoning semantics to label statements as accepted, rejected, or ambiguous, depending on the argumentative structure and rule superiority relations.

DAMN also features a REST API that enables external integration, and its modular architecture supports semantic parameterization such as toggling between ambiguity blocking versus propagating and enabling team defeat. The architecture is designed to support collaborative workflows, allowing multiple agents to input preferences and justifications in parallel. This makes DAMN especially well-suited for online participatory platforms and multistakeholder decision contexts.

DAMN use case: ArgCW-LCA

The growing emphasis on sustainability in both public and private sectors has made life cycle assessment (LCA) a widely used tool for evaluating the environmental impacts of production and consumption systems. LCA aggregates impacts across criteria such as ecosystem damage and human health into either a set of midpoints or a single score (endpoint) to support decision-making. However, the process of weighting these criteria is inherently subjective and may introduce biases that reduce the acceptability of decisions. This can ultimately lead to reduced public trust or environmental inequalities.

To address these concerns, the ArgCW-LCA methodology (presented in Sohn et al., 2020) was developed as a transparent and inclusive methodology to support LCA-based decisions. It enables practitioners to integrate diverse stakeholder perspectives in the development of weightings, thereby reducing bias and increasing transparency in interpretation. This approach supports more legitimate and well-supported decisions, particularly valuable in complex sustainability challenges such as biowaste valorization.

42. Note that DAMN is both the name of the platform and its visualization module.

43. <https://graphik-team.github.io/graal/>

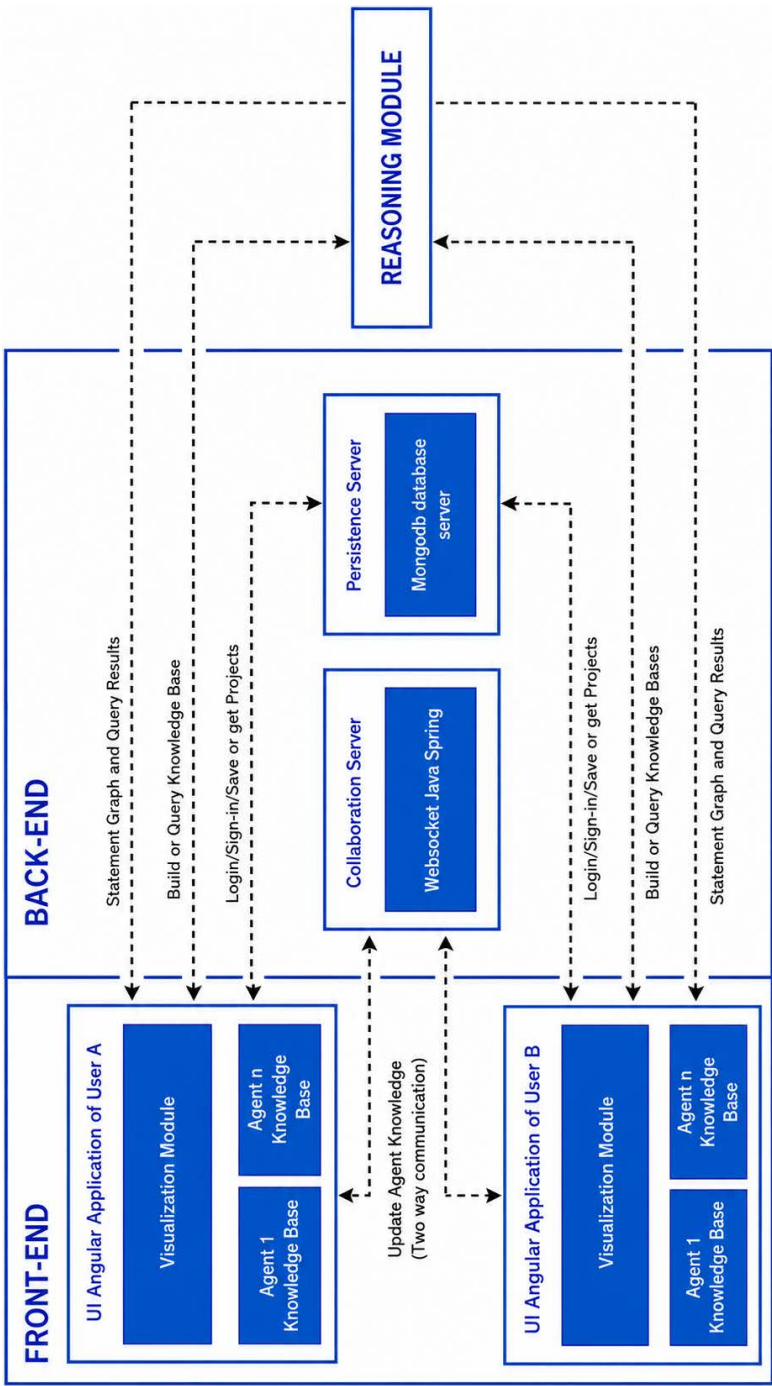


Figure 5.4. DAMN general architecture.

ArgCW-LCA was developed within the context of the European H2020 project No Agricultural Waste (NoAW) that was concerned with the lost potential of agricultural waste. Indeed, agricultural residues represent a huge pool of untapped biomass (up to 50% of fresh harvested crops) that may even become an economic and environmental burden. Driven by a near zero-waste goal and supported by 32 transdisciplinary academic and private partners from Europe and China, NoAW aimed to convert agricultural waste into eco-efficient bioenergy and biobased products. The project adopted a holistic life cycle perspective, involving regional and seasonal considerations, and promoted the inclusion of all stakeholders in the agricultural chain. In this setting, ArgCW-LCA was developed to help assess the sustainability and decision pathways for valorizing such waste using structured and inclusive argumentation-based methods.

In order to do so, ArgCW-LCA uses DAMN to represent, and compute interactions between, stakeholders' arguments about the relative importance of LCA criteria. For instance, Figure 5.5 presents a small excerpt in which several actors put forward arguments concerning the LCA criterion "water consumption." At the bottom of the figure, one actor justifies assigning a very high importance to this criterion (*rank(waterConsumption, 1)*) based on its global impact on water resources. This argument is, however, challenged by another actor, who argues that water consumption only has a local impact on water resources, making the conclusion ambiguous (in orange). In the central part of the figure, an actor assigns a medium level of importance to the criterion (*rank(waterConsumption, 3)*), arguing that solutions exist to mitigate the impact of water consumption; this argument is not attacked by any counterargument and is therefore accepted (in green). Finally, at the top of the figure, an actor puts forward an argument that is deemed logically invalid and is therefore automatically rejected by the system (in red). From these collectively justified importances, the system computes a stakeholder-informed weighting profile that feeds into a multicriteria decision analysis tailored for LCA in biowaste valorization.

The ArgCW-LCA methodology refines DAMN's general workflow as follows:

- Stakeholder input: collect stakeholder perspectives on the relative importance of various LCA criteria
- Preference conversion: convert these perspectives into quantitative weights
- Accord assessment: evaluate the level of agreement among stakeholders to identify controversial criteria
- Argumentation handling: use argumentation techniques to address and reconcile divergent perspectives on these criteria by identifying accepted and rejected arguments
- Weight calibration: calculate and calibrate final weights based on argumentation results
- Decision application: apply the calibrated weights on the criteria of an LCA methodology to assess and compare valorization routes

The impact of the ArgCW-LCA is shown in Section 5.3.

5.2.2 MyChoice

Presentation of MyChoice

MyChoice (Thomopoulos et al., 2020) is a user-friendly web-based application supporting collective decision-making developed by the French National Research Institute for Agriculture, Food and Environment (INRAE). It is designed to use

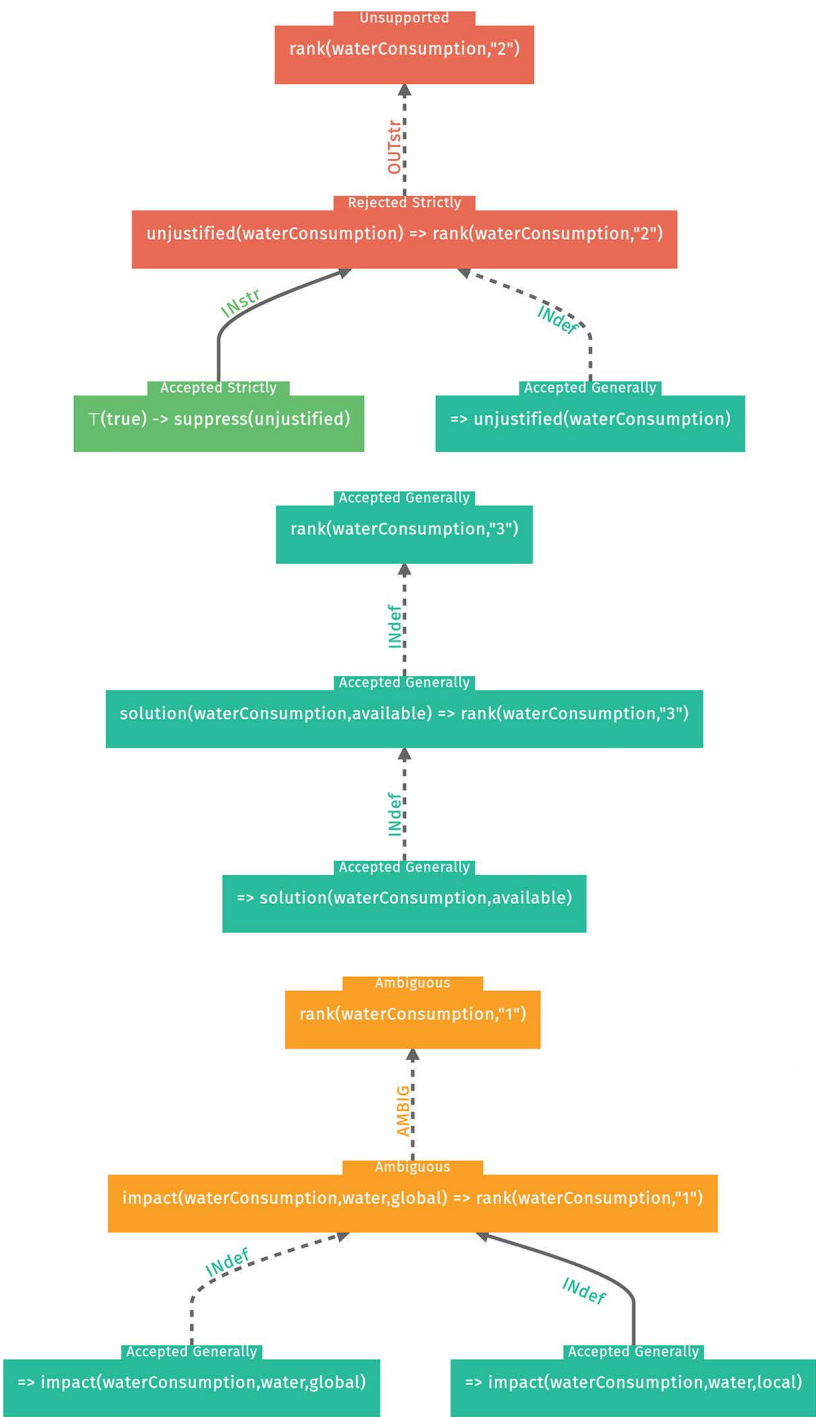


Figure 5.5. An illustrative example of DAMN in the context of the ArgCW-LCA methodology showing statements supporting or rejecting specific rankings for the “water consumption” LCA criterion.

explanatory arguments from multiple stakeholders and sources to analyze, compare, and assess the acceptability of various alternatives based on different criteria and aims. Because it can be used for such applications as technologies, food processes, or product variants, it is well-suited for work dealing with new trends and developments in food chains that require various stakeholders to cooperate and reach a consensus. That said, the tool's design is generic and may also be applied to other fields. The design concepts of MyChoice come from different disciplines, including multicriteria decision-making, AI argumentation, database information systems, and social psychology. The MyChoice tool is based on the analysis of arguments expressing the knowledge and opinions of different types of stakeholders concerning the advantages (expressed by "pro" arguments) and disadvantages (expressed by "con" arguments) of a number of alternatives considered for different criteria. It aims to reflect stakeholders' attitudes regarding these alternatives, to measure the acceptability of the alternatives for stakeholders, and to support collective decision-making. Alternatives could include choices such as animal-based versus plant-based diets, white versus wholegrain bread, high-speed detection technologies versus classic ones, to name just a few. The combination of both "pro" and "con" arguments falls under the category of "bipolar" argumentation. Some specificities of the MyChoice approach are:

- The multicriteria as well as multistakeholder approach to structuring arguments.

The following features are common with the multicriteria decision framework (Belton and Stewart, 2010): Alternatives, Criteria, and Aims, which can be considered as refinements of criteria, and Attributes (called "Properties" in MyChoice). Arguments are also organized according to the stakeholders concerned.

- In MyChoice, "argument" means the expression of an opinion by a stakeholder, in favor of or against an alternative, pursuing a particular aim (e.g., consuming a balanced diet) itself linked to a more general criterion (e.g., nutritional, sensory, economic, social, ethical). This opinion is justified by a property of the alternative that is supported or rejected (e.g., price, protein content), which is declared to take on a certain value (e.g., high, balanced). An argument in MyChoice can thereby be defined as a tuple composed of an ID, a stakeholder who expressed the argument, an alternative considered, type of argument (pro or con), etc. The detailed structure is represented through a database schema.

- The definition of how a "collective attitude" can be computed, with regard to each alternative evaluated. In the expectancy-value theory (Fishbein, 1967), which is a general model of human decision-making, individuals are assumed to make the choice associated with the most desirable outcomes (i.e., the one evaluated most positively). This global evaluation, denoted as "attitude," is derived from the perceived likelihood that the choice is associated with a number of key outcomes, weighted by the evaluation of those outcomes. Formally, the attitude is computed as follows:

$$\text{Attitude} = \sum_{i=1}^n b_i e_i$$

where b refers to the outcome belief, e refers to the evaluation of that belief, and n is the number of criteria. MyChoice introduces a practical way of computing b and e , in collective cases:

- e_i is computed as the proportion of arguments claiming the considered alternative allows the aim i to be reached within all arguments (pro and con) concerning the achievement of aim i for the considered alternative.

- b_i is computed as the proportion of arguments related with aim i in the whole debate, i.e., for all alternatives.

The MyChoice workflow (shown in Figure 5.6) is as follows:

- Argument collection: gather arguments from diverse stakeholders concerning various scenarios (e.g., insect-based biowaste valorization)
- Contextual analysis based on arguments' metadata: assess arguments based on metadata, such as authorship, purpose, and alignment with project goals
- Visualization and evaluation: present aggregated insights and calculate stakeholder attitudes to identify areas of agreement and contention

MyChoice focuses on highlighting the external properties of arguments, such as their source and objectives, to provide a holistic view. Its key strengths are:

- Quantitative assessment of arguments and stakeholder perspectives
- Strong emphasis on visualizing trends and conflicts among stakeholders

MyChoice implementation

The MyChoice software is a web application based on a client-server architecture. The client user interface is developed in JavaScript with Vue.js. It relies on services provided by the server, which is implemented in Java. The exchanges between the client side and the server side occur through the HTTP protocol, using JSON as an exchange format.

The input is a relational database conforming to the MyChoice schema. Alternatively, a spreadsheet format as well as a CSV format are provided for data inputs. An example of argument description in spreadsheet format for the use case data is shown in Figure 5.7.

The outputs are different ways of synthetically visualizing the arguments and their distribution in criteria and stakeholder categories, as illustrated in Figure 5.8; color-code analysis; collective attitude computation; and different modes to filter and analyze the arguments (consensus, expertise, data reliability, etc.).

MyChoice use case: insect-based biowaste conversion

In the use case considered (Vivas, 2022), several scenarios of insect-based biowaste conversion – a promising technology for the valorization of biowaste through insect breeding – are explored and compared at an early stage based on available knowledge. Insect-based conversion consists in raising insects as an ecological way of recycling and recovering biowaste (e.g., in the form of protein for animal feed, fertilizer, biofuel, etc.). In the urban and peri-urban environment, where waste management is a major issue, insect-based conversion offers an attractive solution but also raises specific questions, particularly in terms of biowaste heterogeneity, management of safety risks, and regulatory and societal constraints. Several studies have dealt with evaluating the potential of this solution at an early stage, by anticipating both the benefits and risks of insect-based conversion of different types of substrates and especially biowaste, such as the FLY4WASTE project that the Vivas (2022) study belongs to, following the creation of the INSECT4CITY consortium of 17 research units (Engel et al., 2022).

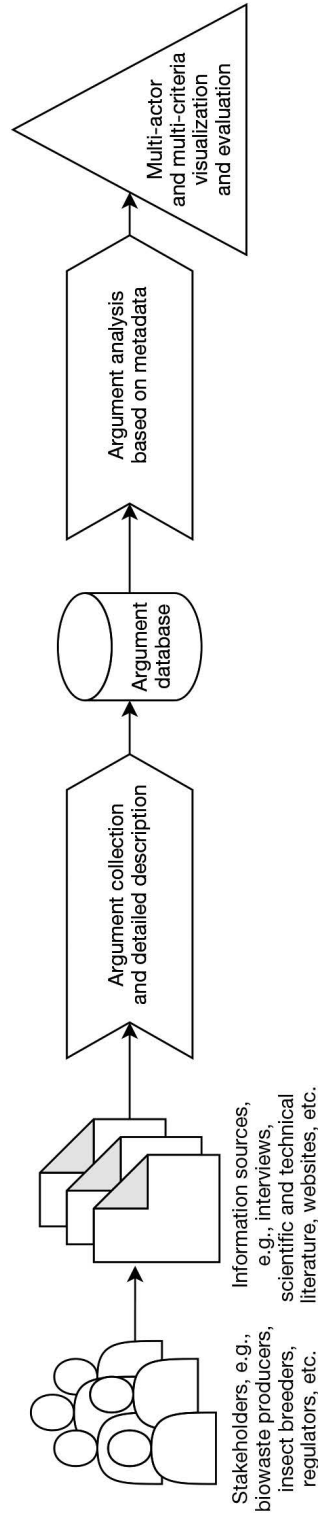


Figure 5.6. Illustration of the MyChoice workflow.

id	Arg name	Stakeholder	name	Alternative	type	name	Criterion	aim	name	Property	value	assertion	explanation	date	name	Source
1	Insect breeders	Animal feed +	+	+	Microbiological	Preventing microbiological contamination in finished pro	Effective	The analyses	When an analysis	23/03/2022	Entofood (Malaysia)					
2	Insect breeders	Animal feed +	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Effective	The analyses	When an analysis	23/03/2022	Entofood (Malaysia)					
3	Insect breeders	Biofuel + fertiliz	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Effective	The analyses	When an analysis	23/03/2022	Entofood (Malaysia)					
4	Insect breeders	Animal feed +	+	+	Microbiological	Preventing microbiological contamination in finished pro	Effective	The analyses	When an analysis	2021	Industrialist in the int					
5	Insect breeders	Animal feed +	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Effective	The analyses	When an analysis	2021	Industrialist in the int					
6	Insect breeders	Biofuel + fertiliz	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Effective	The analyses	When an analysis	2021	Industrialist in the int					
7	Insect breeders	Animal feed +	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2020	The Core Gut Micro					
8	Insect breeders	Biofuel + fertiliz	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2020	The Core Gut Micro					
9	Insect breeders	Animal feed +	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2019	Assessing the Micro					
10	Insect breeders	Biofuel + fertiliz	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2019	Assessing the Micro					
11	Insect breeders	Animal feed +	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2018	Hermelia illucens (D					
12	Insect breeders	Biofuel + fertiliz	+	+	Microbiological	Avoid microbiological contamination that impacts insect	Adaptive	Microbiota	Not all microot The larvae are abl	2018	Hermelia illucens (D					
13	Insect breeders	Animal feed +	-	-	Microbiological	Preventing microbiological contamination in finished pro	Raised	Microbiological contain	The products that	2022	Symposium New St					
14	Insect breeders	Animal feed +	-	-	Microbiological	Avoid microbiological contamination that impacts insect	Raised	Microbiological contain	The products that	2022	Symposium New St					
15	Insect breeders	Animal feed +	+	+	Microbiological	Preventing microbiological contamination in finished pro	Raised	Microbiological contain	The products that	2018	Food Safety Issues					
16	Insect breeders	Animal feed +	-	-	Microbiological	Avoid microbiological contamination that impacts insect	Raised	Microbiological contain	The products that	2018	Food Safety Issues					
17	Insect breeders	Animal feed +	-	-	Microbiological	Preventing microbiological contamination in finished pro	Raised	Microbiological contain	The products that	2021	Industrialist in the int					
18	Insect breeders	Animal feed +	-	-	Microbiological	Avoid microbiological contamination that impacts insect	Raised	Microbiological contain	The products that	2021	Industrialist in the int					
19	Insect breeders	Animal feed +	+	+	Microbiological	Preventing microbiological contamination in finished pro	Raised	Microbiological contain	The products that	2021	Researcher					
20	Insect breeders	Animal feed +	-	-	Microbiological	Avoid microbiological contamination that impacts insect	Raised	Microbiological contain	The products that	2021	Researcher					

Figure 5.7. An extract of MyChoice input data in spreadsheet format.

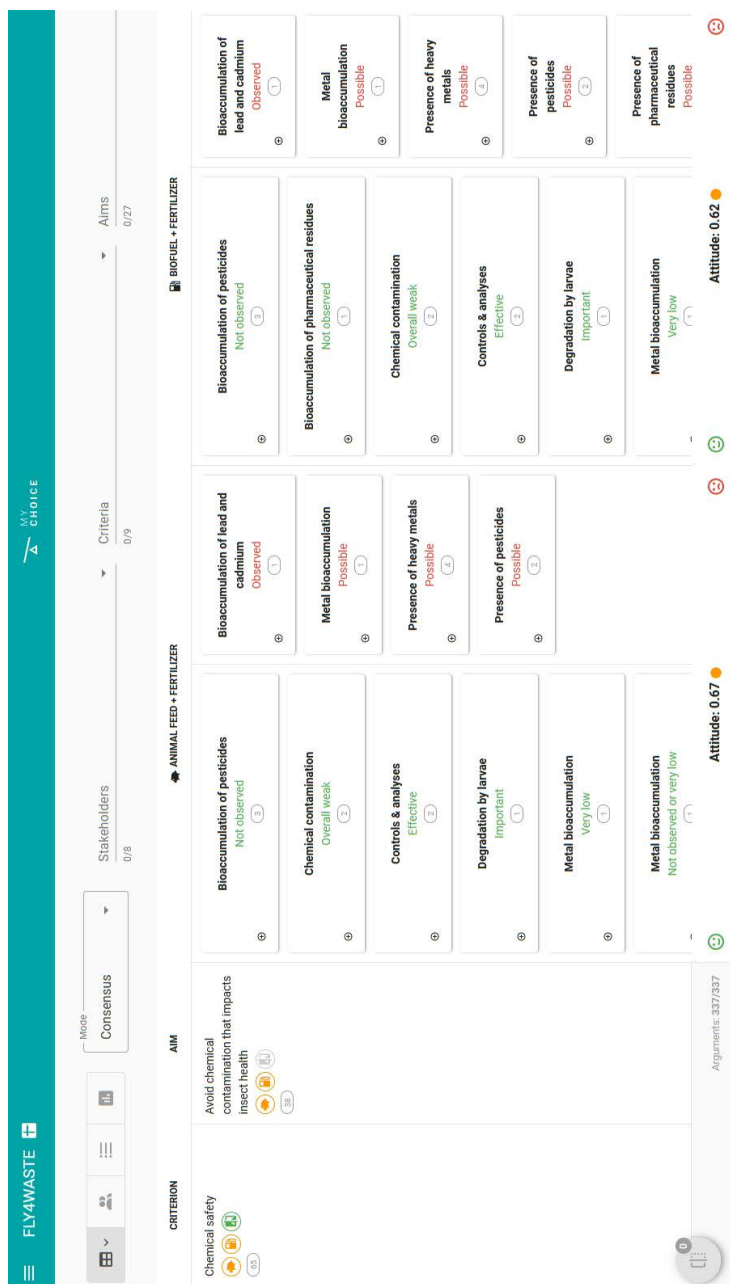


Figure 5.8. Example of MyChoice interface presenting the arguments per criterion and aim (first two columns), distinguishing between pro arguments (green labels) and con arguments (red labels) for each alternative (“Animal feed + fertilizer” versus “Biofuel + fertilizer”).

The use case presented focuses on this risk–benefit analysis. It aims to answer the following questions in order to assess different insect-based conversion scenarios differing in their input and output materials:

- What is the overall attractiveness of each scenario? Can the scenarios be ranked?
- What is the representation, in terms of the amount of information available, of each criterion considered (health, nutritional, etc.) for the different scenarios? Of each stakeholder involved?
- How attractive is each scenario from the perspective of each criterion? Of each stakeholder? Why?
- Can we compare scenarios by criterion? By stakeholder?
- Which stakeholders are most satisfied by which criteria, in each scenario?

A prerequisite for the effective and relevant identification of available information is to precisely delimit the system studied. Here, we are dealing with scenarios defined by the substrate used, the targeted valorization product, and the insect considered. The study was limited to one insect, the black soldier fly (*Hermetia illucens*), and to the following three scenarios:

- Scenario 1: “Animal feed + fertilizer.” Substrate: agrifood by-products (e.g., residues, noncompliant products). Valorization product: insect larvae for animal feed and frass as fertilizer, which can be used to make organic fertilizers. Insect: *Hermetia illucens*.
- Scenario 2: “Biofuels + fertilizer.” Substrate: compostable urban biowaste (household and restaurant waste). Valorization product: biofuel from oil extracted from insects, and frass as fertilizer. Insect: *Hermetia illucens*.
- Scenario 3: “Business-as-usual.” This scenario considers the case where no innovative solutions are implemented. It serves as a “control” in the scenario analysis.

Limiting the study to these three scenarios considerably reduces the sources of information available to analyze this emerging sector and brings us close to exhaustiveness. The types of sources used include (i) peer-reviewed scientific articles, (ii) technical reports and information published on websites, and (iii) scientific and industrial conferences and research project meetings.

Part of this information comes from the 14 webinars organized through the INSECT-4CITY consortium, in which 18 researchers and 4 industry stakeholders participated, and which covered various topics: insect biology, conversion processes, biowaste recycling, animal and human nutrition, chemical and microbiological safety, consumer acceptability, multicriteria evaluation, and environmental, economic and social sustainability. These webinars and the information shared within the network made it possible to constitute an initial multidisciplinary bibliography. Subsequently, using the online scientific documentation tool “Connected Papers,” an expanded bibliography was created, supplemented by queries of databases of scientific resources with contractual access (Web of Science, Scopus, EconLit, Food Science Source, TAIR, among others) or open access (HAL, ORBi, etc.), as well as Google Scholar, using the descriptive keywords of the scenarios. Finally, the “New Stakes and Challenges of the Industrial Insect Rearing Industry” conference, held in April 2022 in the Paris region (Adebiotech⁴⁴), made it possible to supplement the project data with industrial sources.

44. New stakes and challenges of the industrial insect rearing industry. Retrieved from <https://asso.adebiotech.org/new-stakes-and-challenges-of-the-industrial-insect-rearing-industry/>

► 5.3 Outcomes

5.3.1 DAMN

Using the ArgCW-LCA methodology, DAMN was applied within the Horizon 2020 NoAW project on a dataset composed of preferences and justifications from 31 experts (both public and private stakeholders). The resulting argument-based criteria weighting was then fed into a multicriteria analysis method (TOPSIS). In addition to the practical application within the NoAW project, the ArgCW-LCA methodology demonstrated its capacity to generate meaningful stakeholder engagement and differentiated decision outcomes. In terms of LCA practice, ArgCW-LCA enables specialists to:

- Account for the full range of environmental impacts evaluated in LCAs
- Include nonenvironmental dimensions such as economic and societal impacts
- Derive transparent weightings from stakeholder preferences
- Incorporate stakeholder viewpoints through justification analysis

The method was applied to a real-world decision in the NoAW project, where stakeholders had to select a polyphenol extraction technology for upscaling. By integrating structured argumentation, ArgCW-LCA helped clarify the rationale behind weighting decisions and exposed implicit trade-offs. Notably, the results demonstrated that applying stakeholder-informed weightings, especially when filtered through justification analysis, could significantly affect final rankings, emphasizing the role of inclusive reasoning in sustainability assessments.

ArgCW-LCA was applied to derive stakeholder-informed weights for LCA criteria, used to rank four biowaste valorization technologies (polyphenol extraction methods). The innovation lies in using argumentation (via DAMN) to assess and calibrate these weights, rather than relying on simple preference aggregation. The stakeholders provided rankings and justifications for a set of LCA impact categories from the ReCiPe 2016 methodology (Huijbregts et al., 2017), including, for instance, global warming potential, water consumption, land use, and human toxicity.

The analysis of stakeholder preferences and their associated justifications revealed significant disagreement. For instance, some stakeholders prioritized global warming due to long-term global consequences, others emphasized water use based on regional scarcity concerns, and a few preferred land use due to its perceived socioeconomic relevance. Some criteria, such as fossil resource scarcity, were perceived as less important and there was less disagreement about them.

By applying DAMN's reasoning to analyze the validity and clarity of justifications, ArgCW-LCA filtered out unjustified rankings (arguments that are either not supported by facts or by another accepted argument) or weakly argued rankings (those that are "beaten" by stronger arguments). This had two major consequences. First, the final weighting profile, after applying argumentation, significantly diverged from the original (unfiltered) profile, as can be seen in Table 5.1. In particular, criteria with many unsupported or weakly justified rankings (e.g., abiotic depletion) received lower weights, and criteria with consistently strong justifications across agents (e.g., global warming and water use, for which there was relative consensus) were upweighted. Second, as a consequence, the TOPSIS ranking of the four polyphenol extraction technologies

changed depending on whether original or argumentation-corrected weights were used, illustrating the importance of how justifications about criteria importance can alter decision support results (for more details, see Sohn et al., 2020). A major contribution of this work is to provide a justification, and traceability, for the criteria weights.

Table 5.1 Example of modifying the LCA criteria relative importance after applying the ArgCW-LCA methodology. The greater the weight, the greater the importance. Adapted from Sohn et al. (2020).

Criteria	Fine particulate matter formation	Fossil resource scarcity	Fresh-water eco-toxicity	Freshwater eutrophication	Global warming	Human carcinogenic toxicity	Human non-carcinogenic toxicity	Ionizing radiation	Land use
Original	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5
ArgCW- LCA	17.87	78.83	204.59	100.58	64.53	95.00	3.28	139.00	17.36

Criteria	Marine ecotoxicity	Marine eutrophication	Mineral resource scarcity	Ozone formation, human health	Ozone formation, terrestrial ecosystems	Stratospheric ozone depletion	Terrestrial acidification	Terrestrial eco-toxicity	Water consumption
Original	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5	47.5
ArgCW- LCA	101.09	1.69	0.00	30.65	36.86	2.49	32.38	57.25	16.52

5.3.2 MyChoice

From the information collected, arguments were extracted and modeled according to the MyChoice structure (Vivas, 2022). In all, 337 arguments were obtained and stored: 178 concerned Scenario 1, 111 Scenario 2, and 48 Scenario 3. Analyzing these arguments through the MyChoice approach:

- Provided quantitative measures of stakeholder attitudes toward insect-based biowaste conversion scenarios
- Enhanced understanding of social acceptability and economic feasibility through detailed visualization of argument trends
- Delivered insights by balancing diverse perspectives

Specific use-case results

Overall attractiveness and ranking of the scenarios. The collective attitude was computed for each scenario, according to the arguments collected. The results reflected a positive attitude for both insect-based conversion scenarios (although it was more marked for Scenario 1 – Animal feed), and a negative attitude for Scenario 3 – Business-as-usual.

Number of arguments identified by criterion or by stakeholder. The number of arguments reflects the amount of information available in documents and debates regarding

the criterion/stakeholder considered, and thus the focus it has at a given moment in time. The quantitative distribution of arguments according to the different criteria reflected different concerns between Scenarios 1 and 2. For Scenario 1, the majority of the arguments concerned chemical safety (23% of arguments), economic profitability (17%), microbiological safety (15%), environmental impact (12%), nutritional interest (11%), and social acceptability (10%), with the other criteria representing less than 10% of the arguments. For Scenario 2 – Biofuel, economic profitability was the top concern (21% of arguments) while nutritional aspects were limited by nonfood use (biofuels). The stakeholders involved were more varied in Scenario 1 (insect breeders, but also citizens, other breeders, biowaste suppliers, etc.).

Attractiveness of scenarios by criterion or by stakeholder. In Scenario 1, the criteria linked to regulation, nutritional interest, and environmental impact received a very positive assessment. At a finer level, examination of aims provided reasons for this outcome. For example, the chemical safety criterion covers three types of concerns: avoiding chemical risks in frass used as fertilizers (very favorable opinion), avoiding chemical contamination that impacts insect health (mixed opinion), and avoiding chemical contamination in finished products (mixed opinion). The same analysis can also be performed by stakeholders.

Figure 5.9 shows an example of visual representation illustrating MyChoice analysis per criterion, in the case of Scenario 1. Green circles represent positive arguments, whereas red circles represent negative ones.

5.3.3 Comparison

DAMN in the argumentation landscape

Beyond the two main case studies, it is helpful to understand how DAMN compares to other argumentation tools found in the literature. Tools such as ASPIC⁺, DeLP, Carneades, and ArguMed offer powerful capabilities for formal reasoning, but often lack some features that make DAMN particularly suited for collaborative decision-making:

- Web-based interface: Unlike many desktop-only or script-driven systems, DAMN provides an accessible online platform, allowing diverse stakeholders to contribute arguments asynchronously.

- Agent collaboration: DAMN supports multi-agent reasoning, letting different users (agents) input conflicting preferences and justifications. These are resolved through defeasible logic, enabling debate-like collective decision.

- Visualization: The Statement Graph model is presented as an interactive graph, where support/attack relations are visibly rendered. This enhances comprehension and traceability.

- Customizable semantics: Unlike other tools, DAMN allows users to select between ambiguity blocking or propagating, and enables team defeat reasoning. These options provide greater flexibility in the modeling of real-world uncertainty, compromise, and collaboration.

In contrast, tools like ASPIC⁺ focus on formal proof structures and logical rigor but typically require more technical input formats and lack web-based usability or real-time feedback. Carneades and ArguMed offer user-friendly interfaces, but feature limited support for distributed reasoning or detailed justification analysis.

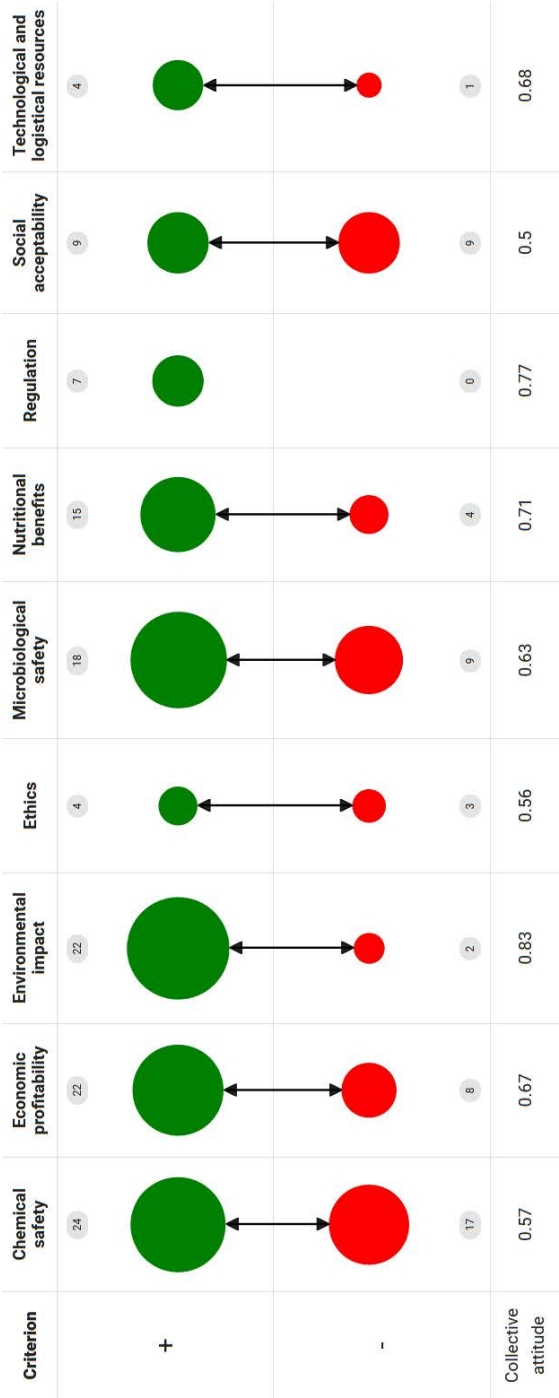


Figure 5.9. Example of MyChoice visual analysis per criterion, in the case of Scenario 1.

MyChoice in the argumentation landscape

The MyChoice model originally stems from the seminal theoretical works on argumentation-based decision support (Amgoud and Prade, 2009). Yet it took several liberties compared to its predecessors, going so far as to provide an implementation. The major differences are a detailed description of argument metadata allowed by a database representation of arguments, which to our knowledge is unique; the introduction of various analysis modalities and visualizations; and a concrete aggregation method based on the definition of collective attitude.

Regarding MyChoice, the main features characterizing its specific positioning are:

- its user-friendly design and interface, aimed at nonspecialists;
- its focus on the context of arguments through detailed metadata (who speaks, with what objective, in what source, etc.) to provide a holistic view;
- its “collective attitude” concept and computation, inspired by social psychology.

Key similarities and differences between DAMN and MyChoice

Table 5.2 summarizes the key similarities and differences between DAMN and MyChoice.

Table 5.2 Similarities and differences between the DAMN and MyChoice tools

Feature	DAMN	MyChoice
Focus	Argument content and logical structure	Argument metadata (e.g., author, aim, criterion) and bipolar status of arguments (pro or con)
Argumentation approach	Structured argumentation via Statement Graphs	Decision-oriented argumentation framework
Semantic customization	Ambiguity blocking/ propagating semantics, and team defeat semantics	Various analysis modes (consensus, expert, data reliability, stakeholder/criterion, etc.) based on specific selections in the argument database
Collaboration	Multi-agent, supports conflicting justifications and collaborative use	Multistakeholder, supports conflicting viewpoints, possible collaborative online data input
Visualization	Statement Graphs	Tabular/global argument-attack graphs/charts
Evaluation output	Argument status	Collective attitude score

Both approaches begin by gathering and structuring arguments, which forms the foundation for their respective analyses. In each case, arguments are collected from stakeholders, organized according to the method’s framework, and prepared for evaluation.

They each employ systematic frameworks to assess and use these arguments in decision support, but their focus diverges notably. DAMN concentrates on the qualitative content and logical structure of arguments, modeling their internal reasoning and the relationships between claims. This makes it especially suitable for contexts where the depth, validity, and interconnections of arguments are crucial to the decision-making process. MyChoice, on the other hand, emphasizes evaluating arguments through their metadata – such as the source, intended purpose, associated criteria, and even the number of times they occur – allowing it to detect patterns, trends, and stakeholder dynamics without requiring full formalization of content.

As a result, DAMN is particularly well aligned with technical and content-rich analyses where the quality of reasoning is central, while MyChoice excels in participatory scenarios where capturing the breadth of perspectives, visualizing consensus and dissent, and facilitating stakeholder engagement are primary goals.

► 5.4 Conclusion and resources

5.4.1 Future directions for DAMN

DAMN has shown a solid ability in deep, structured argumentation, especially where justifications are crucial to unpack preference complexity. Future enhancements could focus on lowering the barrier to entry by developing natural language interfaces or structured templates to guide users in expressing justifications. Currently, inputting formalized statements requires some technical fluency. Bridging this gap would allow nonexpert stakeholders to contribute meaningfully, further democratizing decision-making.

Additionally, DAMN could be integrated into participatory digital platforms, such as stakeholder deliberation spaces or policy co-design workshops. These platforms could support asynchronous input, live discussion around contested justifications, and visualization of evolving consensus or disagreement. As food systems grow increasingly complex, such integrations could extend DAMN's utility beyond expert panels into broader societal contexts.

5.4.2 Future directions for MyChoice

By using metadata to frame argument relevance and stakeholder attitudes, MyChoice is already relevant to participatory contexts. One direction for future development is to enrich the evaluation layer, potentially combining metadata-based evaluation with simple semantic parsing to check coherence or alignment with goals. Improvements could also include filtering of argument quality and enhanced feedback mechanisms to guide participants in formulating stronger, more targeted arguments.

Another promising avenue lies in scenario replay and participatory forecasting: using MyChoice to compare evolving positions across time or simulate the effects of new stakeholder inputs.

5.4.3 Synergies and hybrid approaches

The comparative analysis of DAMN and MyChoice highlights the diverse capabilities of argumentation-based decision support systems. While DAMN provides detailed insights by focusing on the content of arguments, MyChoice offers a higher-level approach by emphasizing aspects based on arguments' metadata and visualizing trends. However, while both tools operate on different facets of argumentation, they are not mutually exclusive but complementary. DAMN provides depth and content-based reasoning, while MyChoice brings accessibility, visualization, and stakeholder engagement. Hybrid approaches could leverage DAMN's rigorous reasoning to assess internally consistent arguments, while MyChoice highlights trends and facilitates broad deliberation, as well as easier access and use.

Combining them could yield systems capable of scaling from local expert input to large-scale participatory processes, an essential capability for decisions in complex food systems. Future research could explore such hybrid models, advancing the field of inclusive and intelligible decision support.

Available resources

Proposed solution	Available resources			
	Main scientific papers	Data papers and datasets	Software codes	Videos
DAMN	Formalism: – Hecham et al., 2018 – Hecham et al., 2020a Tool: – Hecham et al., 2020b – Bisquert et al., 2025 – Bisquert et al., 2024 – Sohn et al., 2020 (ArgCW-LCA)	https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.15454/XVL4BA	https://forgemia.inra.fr/damn/damn-frontend https://forgemia.inra.fr/damn/elder-server	None
MyChoice	Tool concepts and design: – Thomopoulos et al., 2020 Applications: – Thomopoulos et al., 2025 – Vivas et al., 2022 – Kurtz and Thomopoulos, 2021 – Chaib et al., 2022 Conference presentations – Sandjong-Sayon et al., 2022 – Abou Jaoudé et al., 2022	Chaib et al., 2023b	<i>MyChoice</i> was first deposited by INRAE in June 2020 to the European Agency for the Protection of Programs (APP) under version 1.0.0 dated March 26, 2020, No. IDDN.FR.001.280002.000.R.P.2020.000.20900. <i>MyChoice</i> is available from the following depository: https://forgemia.inra.fr/mychoice/public It is distributed under license CeCILL-C (L-GPL type).	FoodSim presentation (English): https://mychoice.netlify.app/presentation_mychoice.mp4 JFN use-case presentation (in French): https://nextcloud.inrae.fr/s/pCs8RXPWR8bNeBj Interface (in French): https://nextcloud.inrae.fr/s/6AdGiSxicMjQYzJ Data format (in French): https://nextcloud.inrae.fr/s/9KgncTjTmeedLHg

Chapter 6

Designing Prospective Scenarios at the Scale of a Value Chain

Romy-Lynn Chaib, Catherine Macombe, Rallou Thomopoulos, Erwan Engel

►► 6.1 Introduction

Agrifood chains are complex dynamic systems involving several stakeholders with divergent opinions and objectives. Guaranteeing the food system's long-term sustainability is thus a complicated task. To be able to transition to a better, more stable state that is compatible with the current and future planetary boundaries and to overcome structural barriers in the food system, all stakeholders must be aware of the potential consequences their actions (or inaction) may have on the value chain itself and on its environment.

Innovations are envisioned at both the research and the industrial scales to move toward safer and more sustainable systems while ensuring economic viability. This chapter considers whether such innovations would profit society, which part(s) of it, in what ways, and under what conditions. To answer this question, a scenario-design approach is introduced and used as a preliminary step to collective decision support presented in Chapter 5. This approach is based on three principles:

- Designing diverse possible scenarios of the considered value chain evolution, with and without innovation. This step belongs to scenario-building prospective methods. To facilitate awareness-raising and decision support, we are developing a methodology to anticipate the important potential impacts of changes in the value chain in the future by identifying key concepts and designing possible scenarios depending on the states of these key concepts in the future.

- Ensuring the practical feasibility of the study from a human point of view, in terms of the availability of stakeholders to be involved in a participatory process. To address this aspect, we suggest innovations to the scenario-design method.

- Ensuring the practical feasibility of the study from a technical point of view. This covers two considerations: first, providing the community with relevant vocabularies to facilitate the structuring of the information gathered, and second, providing timesaving heuristics to the scenario method to make such studies realistic in terms of time cost.

Two projects are taken as examples in this chapter: the EU-China H2020 SAFFI project (2020–2024), and the French National Research Agency SENTINEL project (2020–2025).

The SAFFI project (“Safe Foods for Infants in the EU and China”) sought to enhance food safety in the infant food chain through the development and implementation of novel detection tools, hazard control and mitigation strategies, and advanced models for hazard identification and risk ranking. Within this project, a knowledge engineering component specifically addressed the cost–benefit analysis of various scenarios that incorporated these innovations. The goal was to evaluate the feasibility, efficiency, and overall impact of these new tools on the food chain, considering multiple stakeholders’ perspectives, including technology developers, food industry operators, regulatory bodies, and consumers.

The primary impact assessment question of this study was: What are the global impacts of implementing SAFFI’s innovative tools and strategies in the food chain? The study aimed to identify the benefits and drawbacks of these implementations and provide a ranking of the scenarios based on their overall effectiveness and sustainability. In this chapter, we examine how scenarios were designed in that study: the approach adopted was similar to the classic Godet scenario method (Godet, 2008; Godet and Durance, 2011).

The SENTINEL project (“High-throughput screening tools for a reinforced chemical safety surveillance of food”) aims to improve and strengthen the current chemical safety monitoring system along the food value chain by introducing new screening tools. Indeed, both self-monitoring in the industry and safety control by health authorities are demanding high-throughput, sensitive, and inexpensive screening methods to strengthen the monitoring of priority food hazards. The adopted model is for polychlorinated biphenyls (PCBs) in meat. The screening tools are being developed based on the coupling between latest-generation detectors (mass spectrometry, biosensors, electronic noses) and innovative methods in marker search (omics), sample processing (mixing, extraction), and data processing (chemometrics, bioinformatics).

Part of the project is dedicated to exploring the possible inclusion of the tools envisioned in SENTINEL in the future development of the value chain. Therefore, our objective is to model the likely states of the value chain in the future, first as reference states, without the new tools being implemented (Pesonen et al., 2000), then with the inclusion of the new tools, which impacts the value chain (Hachem et al., 2025).

This chapter is concerned with innovations in the classic scenario method, which was required in this study due to both organizational limitations during the COVID-19 pandemic and time considerations (Chaib et al., 2021, 2022). As such, we use prospective methods (Chaib et al., 2021). This goes beyond the scope of participatory modeling: indeed, it requires not only consulting and discussing with the stakeholders (Barré, 2000; Godet and Durance, 2011; Mermet, 2004), but collaborating with them to co-design a plausible future in order to co-decide what would be best for the value chain. We are thus in a context of collaborative modeling as described in Basco-Carrera et al. (2017).

►► 6.2 Proposed solution

6.2.1 Workflows

Figure 6.1 shows the workflows used for the classic (a), adapted (b), and optimized (c) scenario methods, from stakeholders’ inputs to scenarios.

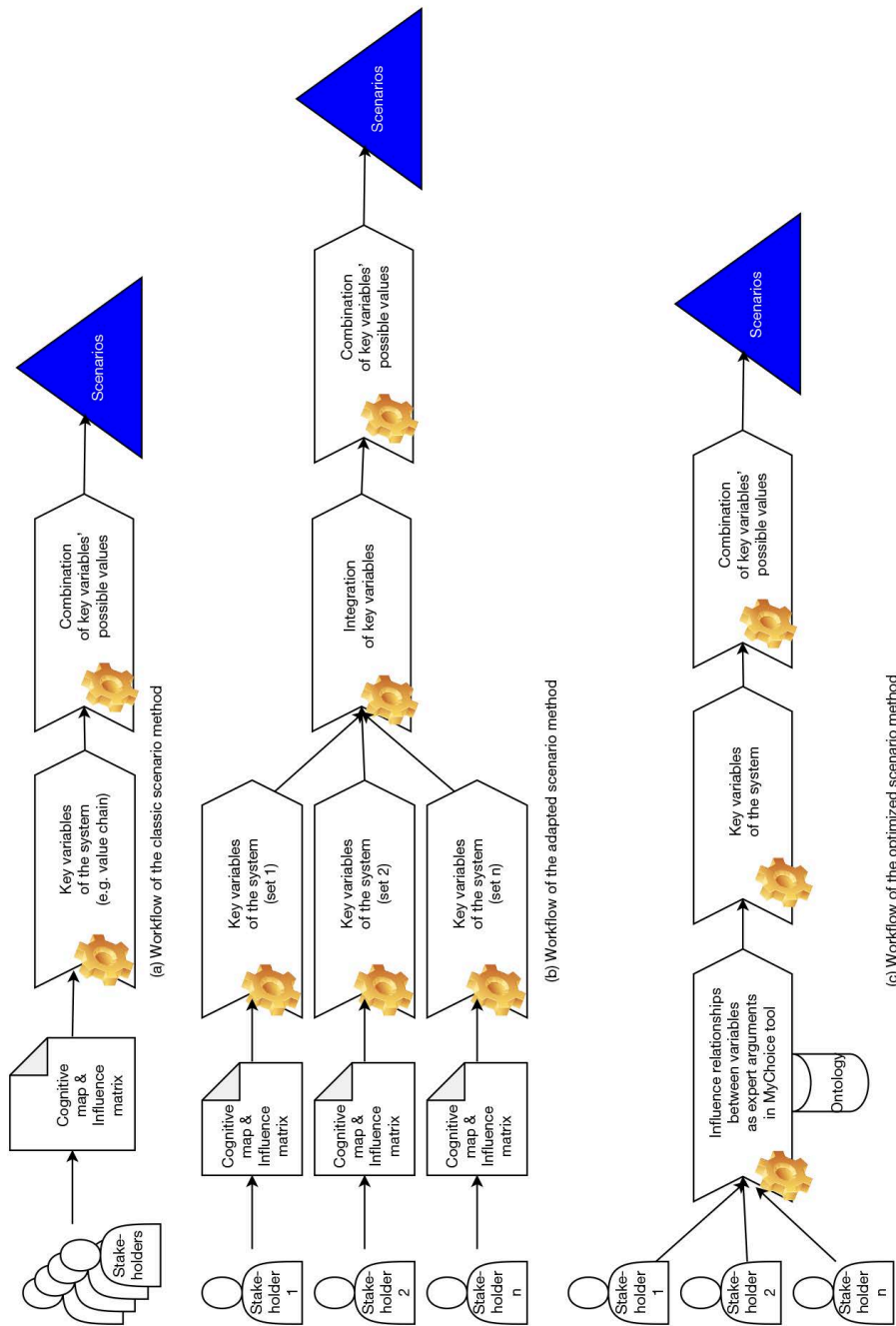


Figure 6.1. Workflows of the three scenario methods.

6.2.2 The classic scenario method

The initial method chosen was the prospective collaborative “scenario method” as described in Godet (2008) and Godet and Durance (2011). By considering possible combinations of variable values, it has the particularity of creating original scenarios that stakeholders may not necessarily think of, which pushes the discussion and results “outside the box.” In this method, the problem of homogenizing ontologies is inexistent since stakeholders themselves meet and establish a consensus on the main ideas to keep in mind.

In the classic scenario method, collective sessions serve to identify the variables and to build consensus about relationships between variables, first in small groups and then by bringing all prospects together. From these group discussions about the relationships between each pair of variables, first cognitive maps followed by matrices of relationships are built for each group. From the consensus built between the different groups, all the relationships (direct and indirect) are summarized in a single matrix, which is then discussed by all prospects, who have the final decision concerning the determination of the meaningful relationships. This whole process is called “structural analysis.” It is the input for determining the system’s key variables, whose value combinations determine the scenarios.

This method was used in the SAFFI project. As an example, in the SAFFI case, one of the key variables collectively identified was “Technological progress.” However, in SENTINEL, a more challenging option was developed out of necessity due to the COVID-19 pandemic. The classic Godet method had to be adapted as it was impossible to have all value-chain stakeholders meet together. The question was thus: How do we conduct a remote foresight study while ensuring correct results and consensus among stakeholders? And how do we do so in a reasonable timeframe? To answer the first question, the classic Godet method was adapted, and to answer the second one, the use of multicriteria argumentation for the purpose of foresight studies is explored.

6.2.3 The “adapted” scenario method: adaptation to the remote context

Since this part of the classic method is based essentially on social interactions, going from collective to individual sessions had methodological repercussions. In the adapted method, structural analysis is based on individual semi-directive interviews conducted with experts who presumably have different views on the sector (political, social, economic, technological, environmental, etc.). It is therefore expected that the variables cited as the main determinants of the system evolution will differ from one actor to another. The adapted method is a combination of linguistic and mathematical approaches. This method was used in the SENTINEL project. It consists in analyzing each interview individually and then analyzing all the interviews as a whole. For instance, as in the SAFFI case, one of the key variables identified was “Technological progress,” but it was obtained as a result of merging several heterogeneous variations of this notion among the different experts (e.g., “increased technicity,” “development of new tools,” etc.). This method is described in detail in Chaib (2023) and Chaib et al. (2023a).

6.2.4 The “optimized” scenario method: heuristic approach

In a context where applying the classic Godet method is not possible, and considering the time and effort needed, the prospective approach of the adapted Godet method has

a lot to gain by being complemented with formal approaches from other disciplines such as knowledge engineering. Indeed, the outputs obtained by applying the adapted scenario method can be obtained through multicriteria/multi-actor argument analysis, using the MyChoice method (Thomopoulos et al., 2020), which concomitantly saves time (Chaib et al., 2022).

Two steps are introduced:

1. Creating an ontology to structure the vocabulary used in the prospective approach

Ontologies represent the concepts used as well as the relationships between them. They allow us to structure knowledge by formalizing information and vocabularies (Maedche and Staab, 2000; Nebot and Berlanga, 2009). Using the same words to refer to one concept is essential when dealing with large and complex knowledge resources, such as agrifood value chains. This is why the process of merging verbatims into concepts and concepts into variables is crucial in both the adapted Godet method and the optimized approach using MyChoice. However, with MyChoice, this process is optimized by building an ontology specific to changes occurring in agrifood value chains as the study goes on (Chaib et al., 2023b).

When doing a foresight study, whether using the adapted Godet method or MyChoice, information comes from different value-chain stakeholders. The question asked is, “According to you, what are the main variables that influence the future of the agrifood value chain and those that depend on it?” The answers are mainly in text form since semi-directive interviews are conducted and then transcribed to ensure proper analysis later on. In addition to the interviews, we added documents since remote work was a necessity during the study period given the ongoing health pandemic. Each document read was considered as a completed interview done (Chaib et al., 2021). When it comes to taking a decision about the value chain, it is important to have divergent opinions (i.e., as many different viewpoints as possible), so it is best to include as many types of stakeholders as possible (Mitchell et al., 1997). The stakeholders involved in the study included raw material producers, breeders, processors, distributors, and consumers as well as public and private institutions, researchers, technical centers, etc. It goes without saying that all of them had different opinions as well as divergent priorities (Handayati et al., 2015), whether because of their position and involvement in the value chain, political commitments, affiliations, or life experiences. They also had various different possible ways of saying the same thing: they use different ontologies to refer to the same concepts. Standardizing the vocabulary used (Chaib et al., 2023d) is a first step toward opinion analysis.

2. Using multicriteria/multi-actor argument analysis for prospective purposes

Multicriteria argumentation is a combination of multicriteria approaches and argumentation (Amgoud et al., 2005; Thomopoulos, 2018). This method makes it possible to deal with qualitative or qualitative criteria. It also helps when dealing with uncertain contexts when making a decision. One of its strengths is that it attributes values to consequences of decisions taken. MyChoice is a tool for multicriteria/multi-actor argument analysis. We chose to use it for prospective purposes for several reasons (Thomopoulos, 2018; Thomopoulos et al., 2020), and especially because we can analyze interviews in a rigorous way to extract relevant and precise information concerning the influence relations between concepts. The MyChoice approach is useful in two different ways when doing a foresight study:

1. In the adapted scenario method, the multiple graphs of influences (one per interviewee) are replaced with a single argument table with homogeneous vocabulary in order to simplify the process. This is achieved by representing each influence relation as an argument as follows:

Given an influence relation from concept A (taking value v_A) to concept B (taking value v_B), beneficial to the scenario of evolution of the value chain, a positive argument (“+,” i.e., favorable) is created, expressing

$A=v_A \rightarrow B=v_B \rightarrow$ support to the scenario.

If the expected effect is unfavorable, a negative argument (“-,” i.e., against) is created. The information is entered in the MyChoice as shown in Table 6.2.

One of the advantages of using this method is that it helps to fix the vocabulary used and to identify nuances in interviewees’ words. Most importantly, from the data entered, the matrix of direct and indirect influence and dependence relations between all sets of variables for all interviews (used as an abscissa in the scenario method) can be easily derived.

2. This multicriteria argumentation tool can be used to confirm the key variables obtained using the scenario method. This option is discussed in Chaib (2023).

For both options, time is saved by integrating all interviews (while homogenizing their vocabularies) as a whole. For instance, the key variable “Technological progress” identified in both previous methods also appears in the optimized method as a criterion in the MyChoice tool, grouping several more specific ones, such as “Bettering food traceability” and “Controlling safety risks,” among others.

Table 6.1 and Figure 6.1 summarize the process and steps of creating scenarios using the classic scenario method, the adapted scenario method, and finally the optimized scenario method that makes use of MyChoice.

►► 6.3 Outcomes

6.3.1 Scenario design

For the SENTINEL project, 21 texts were analyzed (12 transcribed interviews and 9 documents). From those interviews, using the adapted Godet method, 12 variables were identified, with 4 as key variables. Two modalities were distinguished for each of those key variables. This resulted in the obtention of 2^4 scenarios. After removing the incompatible modalities and merging similar scenarios together, six reference scenarios remained. They were submitted to the project partners to choose two of those – contrasted and of interest regarding chemical safety monitoring – to study meticulously. The results and description of the scenarios are presented thoroughly in Chaib et al. (2023a, 2023c).

In addition to creating the scenarios, the concept merging process undertaken to obtain the variables resulted in the creation of an ontology (Chaib et al., 2023b, 2023d). Again, the aim was to simplify the time-consuming task of identifying concepts and merging similar ones into variables.

The comparison of the outputs obtained using the adapted Godet method and the MyChoice tool are presented in Chaib (2023).

Table 6.1 From the classic scenario method to the adapted and the optimized scenario methods

Step	The classic scenario method	The adapted scenario method	The optimized scenario method
Delimit the system under study	<i>Researchers:</i> identify the prospects (no specific method used) <i>Researchers:</i> conduct individual and collective interviews with specialists <i>Prospects:</i> provide variables influencing the system's evolution	<i>Researchers:</i> identify the prospects (using the stakeholder salience model (Mitchell et al., 1997)) <i>Researchers:</i> conduct remote individual interviews + analyze existing documents treated as interviews <i>Prospects:</i> provide variables influencing the system's evolution	<i>Researchers:</i> identify the prospects (using the stakeholder salience model (Mitchell et al., 1997)) <i>Researchers:</i> conduct remote individual interviews + analyze existing documents treated as interviews = INPUTS <i>Prospects:</i> provide variables influencing the system's evolution = INPUTS
Determine the key variables	<i>Researchers:</i> – Make a list of the variables quoted by the prospects – Merge the variables standing for the same concept – Organize groups (e.g., 3 groups of 10 prospects) – <i>Prospects:</i> each group of prospects builds a consensus about the relationships between the variables <i>Researchers:</i> – Build the matrix of relationships between variables for each group – Provide a synthesis matrix to be discussed by the group of prospects as a whole – Select the key variables	<i>Researchers:</i> – List subconcepts (verbatims) quoted by the various sources identified – Group similar ideas into concepts – Convert each interview into a cognitive map to visualize influence and dependence relations between concepts identified – Group concepts into variables – Construct partial squared matrices of variables to identify partial influence and dependence direct and indirect (first order) links – Merge partial matrices into a global one by summarizing all partial direct and indirect links – Identify key variables – Confirm key variables by sending a survey to all interviewed stakeholders	<i>Researchers:</i> – List properties (verbatims) in MyChoice and attribute to each property an aim (concept) and a criterion (variable) – Identify explicit influence relations between variables by filling the spreadsheet – Create the matrix of direct and indirect relationships between variables based on the two columns of the MyChoice spreadsheet that concern the identified variables – Select the key variables – Confirm key variables by sending a survey to all interviewed stakeholders

Table 6.1. (next)

Step	The classic scenario method	The adapted scenario method	The optimized scenario method
Create the base scenarios	<i>Prospects:</i> build a consensus about the main modalities for each key variable	<i>Researchers:</i> – Build preliminary scenarios by combining all possible modalities of all key variables identified in the interviews and in the list of concepts	<i>Researchers:</i> – Build preliminary scenarios by combining all possible modalities of all key variables identified in the interviews and in the list of concepts
	<i>Researchers:</i> envision the different possible combinations of modalities	– Discard scenarios with incompatible modalities <i>Prospects:</i> choose reference scenarios to keep	– Discard scenarios with incompatible modalities <i>Prospects:</i> choose reference scenarios to keep

Table 6.2 Structuring the MyChoice data for prospective building

MyChoice Current Titles	Name Stakeholder	Name Alternative	Pro Con Type	Name Property	Aim	Name Criterion	Value	Condition	Explanation	Name Source	
	Name Stakeholder	Name Alternative	Factor Perception	Criteria	Concept	Variable	Value	Influenced Concept	Influences explanation	Name Source	
Infor- mation entered	Stakeholder group	Business as usual	If the factor influences the value chain positively or negatively	Verbatim	Concept	Variable	Value given to the variable	Influenced concept	Variable influenced by the concept	Clarify the influence relation if possible	Interviewed stakeholder

In the SAFFI project, five reference scenarios were designed (Thomopoulos et al., 2025), each showing varying degrees of impact across the selected criteria:

1. **Technological Advances:** New detection and mitigation tools improved food processing efficiency and safety. However, uncertainties remained regarding organoleptic properties and the high cost of new technologies.

2. **Social Impact:** Scenarios generally indicated positive social outcomes, such as the development of new skills and employment opportunities, though economic access remained a challenge for smaller companies.

3. **Informational Benefits:** Data acquisition and model development showed progress, but data accessibility issues persisted, especially for smaller stakeholders.

4. **Regulatory Impact:** Scenarios highlighted a need for adapted legislation to fully leverage new technologies, with current regulations lagging behind innovations.

5. **Safety and Economic Considerations:** While global food safety improved, economic risks were significant due to the high costs associated with new tools and technologies. The economic burden was particularly high for smaller companies.

6.3.2. Scenario analysis

In SAFFI, the scenarios were then analyzed according to the methodology covered in Chapter 5, and ranked based on their collective attitudes ranging from very favorable to unfavorable. The impacts on different stakeholders, including technology developers, food industry operators, regulatory bodies, and consumers, were qualitatively assessed (Thomopoulos et al., 2025). Comparing the attitudes of stakeholders toward the various impacts allowed us to identify possible hotspots on which value-chain stakeholders could act to limit the unwanted consequences of implementing the change while respecting certain limits and preserving the social fabric woven by the value chain.

One of the key findings was that the “Incentives to technical, digital, and regulatory advances” scenario, which grouped together favorable values of key variables with a focus on targeted tools, was ranked first, followed by the equivalent scenario with untargeted tools (“Democratization of untargeted tools”). The “Small structure-inclusive” and “Untargeted tools for research” scenarios came next, with equivalent attitudes, since their impact concerned part of the stakeholders and was generally more restricted. The “Two-speed economy” scenario that led to heterogeneous economic impacts due to expensive tools was ranked last.

In SENTINEL, two contrasted scenarios were selected by the project partners and meticulously detailed using geographic, economic, and technical data, leading to realistic descriptions of possible future scenarios. This allowed for further specialized studies, such as risk assessment – the project’s core topic – in the prospective scenario.

►► 6.4 Conclusion and resources

By combining prospective participatory scenario design and multicriteria/multi-actor argument analysis, the approach presented in this chapter marks a significant advance in the application of scenario design to the agrifood value chain, particularly in its ability to formalize the anticipation of diverse future states of the system. By incorporating participatory, remote, and optimized methods, this framework adapts to

practical constraints such as time limitations and stakeholder availability while also ensuring the inclusion of a wide range of perspectives, ensuring the robustness and relevance of the scenarios.

The combination of traditional scenario-building techniques with modern tools like ontologies and multicriteria/multi-actor argumentation provides a methodology that is both nuanced and scalable for identifying and evaluating the potential impacts of innovations within the food value chain. Importantly, this approach lays the groundwork for a new branch of social life cycle assessment (S-LCA) by considering the broader social, economic, and environmental consequences of innovations in the value chain. It is extensively developed in Chaib (2023). This new branch is expertise-based through the collection of stakeholders’ arguments. It shifts focus from the impacts on individual products or processes to the collective and dynamic systemwide effects, offering insights to assess future transitions.

This method can be further refined and applied across other sectors, contributing to more informed and collaborative decision-making processes for sustainable development. Ultimately, this work paves the way for a more integrated and participatory approach to evaluating sustainability at the systems level, ensuring that all stakeholders are equipped to make decisions that contribute to a more resilient, equitable, and sustainable future.

Available resources

Main scientific papers	Datasets	Videos	PhD thesis
Thomopoulos et al., 2025 Chaib et al., 2021, 2022 Chaib et al., 2023a, 2023c, 2023d	Dataset resource (ontology): https://doi.org/ 10.57745/YHSDWA	Webinar (in French): https://youtu.be/ 3wWYxauC50	Chaib, 2023

Chapter 7

Assessing the Impact of Participatory Interventions at the Individual Level

Reine Barbar, Rallou Thomopoulos

» 7.1 Introduction

Culinary habits play an essential role in all societies, since they are deeply intertwined with social cultural features. Food habits reflect traditions and are present both in towns and in rural areas, at the domestic level, on the street, and in restaurants, in their traditional forms or with modern trends (Zurayk and Ghyda, 2009). The term “tradition” comes from the Latin *tradere*, which essentially means “to transmit” or “to send.” Tradition thus represents a form of transmission or a flow through time of ideas, praxes, habits, methods, etc. Through such continuity, traditions not only maintain a link to the past but also push us toward the future (Cannarella and Piccioni, 2011). The food engineering/science curriculum has several objectives, and is mainly designed to train graduates to be capable of supporting the needs of society and the food transition, developing short and sustainable circuits, and innovating based on these constraints.

There is currently a gap between the existing food engineering/science curriculum and market needs for graduates with skills in culinary sciences for new product development. The educational challenges are thus adapting the teaching and learning schemes to innovate in a sustainable way based on food heritage, culture, societal needs, and territory. A previous educational project tackled the study of culinary precisions in traditional recipes in Lebanon and in the Middle East, but the innovation lever did not integrate participative societal demands (Barbar et al., 2021). With recent advances in the field of food science, new interdisciplinary concepts have evolved from gastronomy. Food is now also viewed from the perspective of personalized nutrition and includes considerations about swallowing and findings from “soft matter” approaches. Scientists and chefs are seeking new ways to customize nutrition as health and wellness gain prominence across the globe (Subhasri et al., 2022). Food has been shown to be a lever to deal with various social issues, such as integrating people at risk of exclusion (Navarro-Dols and González-Pernía, 2020), contributing to more sustainable consumption habits, educating on health, and supporting the development of local communities (Navarro-Dols et al., 2022).

A framework for valorizing local food heritage in different countries while adapting it to the specific needs of vulnerable publics has been set up for the Erasmus+ project TRADINNOVATIONS.⁴⁵ Figure 7.1 shows some examples of such specific needs.

45. <https://erasmus-plus.ec.europa.eu/projects/search/details/2023-1-FR01-KA220-HED-000166321>

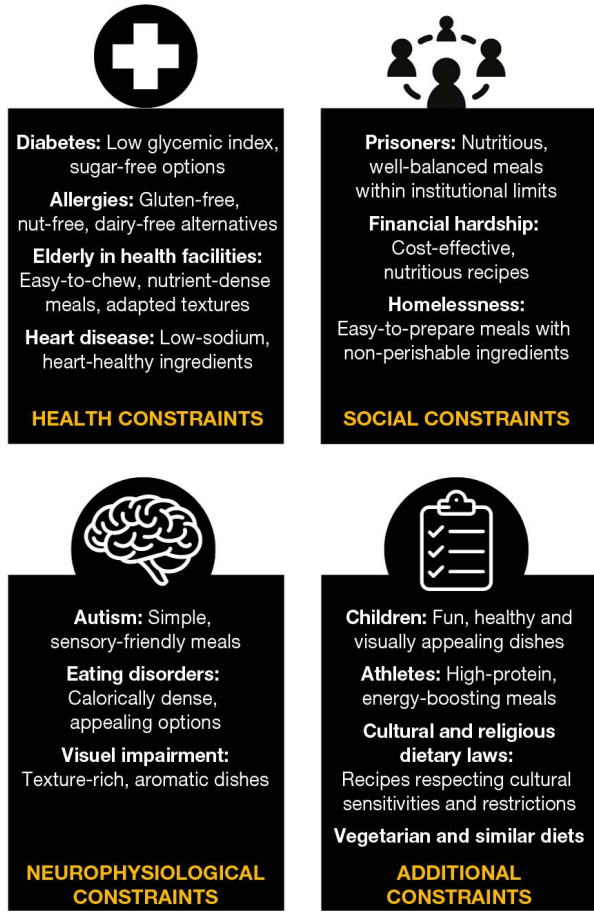


Figure 7.1. Proposition of categories for food constraints for inclusive food innovations in students' TRADINNOVATIONS projects.

The framework first aims to develop an inclusive approach of food toward vulnerable populations, while also proposing an original set of hybrid teaching and learning tools to broaden students' food culture and skills and foster their intercultural and inter-generational interactions. The consortium consists of six academic partners from six European countries. Nine associate partners from four different countries have also joined the consortium. All partners share similar challenges related to education and new food design (at culinary and industrial scales), energy and food transitions as well as cultural and societal diet needs.

In this context, a project-based module entitled "Culinary sciences and societies" in the agricultural engineering curriculum at the Institut Agro-Montpellier offers the opportunity to work each year with a socioeconomic stakeholder with a specific need for inclusive food personalization. This year, we collaborated with the Valentin Haüy association,⁴⁶ which helps people with visual impairments live independently

46. <https://www.avh.asso.fr/fr>

and be fully integrated into society. Indeed, a study by Jones et al. (2019) found that 65% of 101 visually impaired people surveyed found that their visual impairment made cooking difficult. This difficulty often leads to a preference for eating out or consuming ready-made meals, which can have a negative impact on healthy eating behaviors, as Bilyk et al. (2009) discovered. These data highlight the crucial importance of developing recipes to support the culinary autonomy of visually impaired people, enabling them to cook in a safer and more enjoyable way. A multidisciplinary study of ratatouille (a traditional dish from France's Provence region made from fresh vegetables and seasoned with garlic and aromatic herbs) was conducted by students and updated to address target population specifications (Rayaleh and Ang, 2024). Among the limitations of this population, a questionnaire helped to identify the following: food recognition, especially sliced vegetables in sauce (such as those found in traditional ratatouille recipes), difficulties in meal preparation and managing nutritional balance, and accessibility of kitchen tools. A culinary workshop was organized at the Valentin Haüy association's facilities to test the new recipe under real-world conditions. The new recipe, "Three-bite ratatouille," consists of a large savory cake made with layers of sweet bell peppers and zucchini, with a heart of eggplant caviar. This structure creates a unique sensory experience and was chosen to create a balance of textures, with the eggplant caviar acting as a creamier interlude, contrasting with the crunchiness of the vegetables.

The workshop also encouraged personal expression through cooking, linking the act of cooking with memories and emotions, while striking a balance between innovation and tradition to cater for the preferences of all participants, including the oldest among them (Figure 7.2).

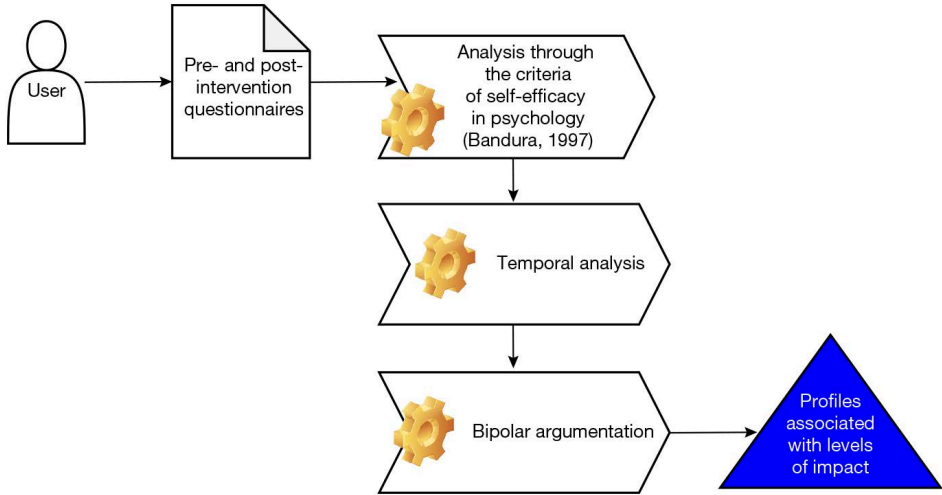


Figure 7.2. Preparation steps during the culinary workshop with participants experiencing visual impairments.

The question dealt with in this chapter is: “How can the impact of an intervention on participants with visual impairments be evaluated?” The study investigates a generic methodology to assess the overall impact of the culinary workshop on a population with specific needs, examining both expected and unexpected outcomes. This includes short-term effects observed during the workshop and potential long-term impacts.

» 7.2 Proposed solution

Figure 7.3 shows the workflow used to assess the impact of an intervention at the individual level.



(a) Workflow to assess the impact of an intervention at the individual level

Figure 7.3. Workflow to assess the impact of an intervention at the individual level.

7.2.1 Participants and setting

The workshop was conducted with visually impaired participants and engineering students specializing in food science. The workshop included cooking sessions with adapted material, and a communal meal to foster interaction and learning.

7.2.2 Data collection

Data were collected using pre- and post-workshop questionnaires administered to visually impaired participants, focusing on the behavioral (habits), cognitive (knowledge), and emotional states of the participants. The pre-workshop questionnaire included six questions about participants’ prior cooking experience and expectations. The post-workshop questionnaire had eight questions about their experience during the workshop, new skills acquired, emotional responses, and future cooking intentions.

7.2.3 Analysis framework

The analysis framework is based on the introduction, within a “bipolar” argumentation approach (i.e., considering both “pro” and “con” arguments, see Chapter 5),

of the psychological theory of self-efficacy by Bandura (1997). Four key factors were considered as analysis criteria:

- Performance accomplishments: Successful experience during the workshop.
- Vicarious experience: Learning by observing peers or instructors.
- Verbal persuasion: Encouragement and feedback received.
- Physiological feedback: Emotional and sensory engagement during the workshop.

Based on participants' responses to the questionnaires, all arguments expressing facilitations or limits to the success of the intervention in relation to the considered objective – gain in self-confidence for autonomous cooking – were classified according to two distinct dimensions.

First, they were organized based on Bandura's four *psychological criteria*: performance accomplishments, vicarious experience, verbal persuasion, and physiological feedback.

Second, these arguments were categorized by their *temporality*, considering whether they occurred before, during, or after the workshop. Counterbalance between pro and con arguments, according to both dimensions, was examined to provide analysis.

►► 7.3 Outcomes

Two main participant profiles emerged based on their objectives and responses.

Participant A: "High potential for improvement" profile (Table 7.1)

- Pre-workshop: This participant had not cooked for five years due to an accident, indicating a low level of current cooking activity but a strong emotional motivation to cook for his daughter.
- During the workshop: Reported positive experiences with cooking techniques, organization and sensory aspects, despite a minor difficulty with using a mandoline.
- Post-workshop: Expressed a strong intention to continue cooking and planned specific future cooking activities.

Participant B: "Already experienced" profile (Table 7.2)

- Pre-workshop: This participant frequently cooked and enjoyed it, indicating a high level of existing cooking activity.
- During the workshop: Reported positive experiences similar to Participant A, including learning new techniques and enjoying the sensory aspects.
- Post-workshop: Indicated no significant change in future cooking activities, reflecting a consolidation rather than a transformation.

►► 7.4 Conclusion and resources

The video of the webinar presenting this work is provided as an openly available resource on: <https://nextcloud.inrae.fr/s/jsPg9x6fAGaTbew> (June 19, 2024, in French).

Table 7.1 Profile of Participant A, showing pro (+) and con (–) arguments for self-efficacy in autonomous cooking, according to the analysis criteria

	Pre-workshop		During the workshop		Post-workshop	
	+	–	+	–	+	–
Performance accomplishments	I used to enjoy cooking	I haven't cooked at all for 5 years, since my accident	Organization/ accessibility: very well adapted	Mandoline: a bit difficult		
Vicarious experience			Shown/ learned: use of a mandoline slicer, whisk, cooking techniques			
Verbal persuasion	I would like to cook for my daughter				I'm going to make a tagine for my daughter next week. I will also buy a mandoline slicer.	
Physiological feedback	Interested		Mandoline: greatly appreciated I really enjoyed it Delighted Had a lot of fun Perception: touch			

Table 7.2 Profile of Participant B, showing pro (+) and con (–) arguments for self-efficacy in autonomous cooking, according to the analysis criteria

	Pre-workshop		During the workshop		Post-workshop	
	+	–	+	–	+	–
Performance accomplishments	I cook often		Workshop experience: very good Practical, very well organized Cutting zucchini, doing it myself			
Vicarious experience			Mandoline: very new Accessibility thanks to volunteers, sighted assistance			

	Pre-workshop		During the workshop		Post-workshop	
	+	-	+	-	+	-
Verbal persuasion						Not much change expected in the future, I already cook a lot
Physiological feedback	I like cooking		Discovery, curiosity, delighted by the atmosphere Touch, taste, presentation: charming, varied Satisfaction, feeling of pride			

Based on the proposed methodology, we can conclude that the workshop had a significant positive impact on participants, particularly those with less prior cooking experience. For Profile A, the workshop was transformative, potentially leading to long-term changes in cooking habits and increased self-efficacy. For Profile B, the workshop reinforced existing skills and confidence.

Further research should consider a more systematic approach to evaluating all stakeholders, including students, volunteers, and organization leaders, to gain a comprehensive understanding of the workshop's broader impacts. Additionally, incorporating feedback from informal exchanges and addressing the diverse preferences in questionnaire design could enhance the evaluation process. Overall, the study highlights the value of tailored, interactive workshops in improving self-efficacy and practical skills among visually impaired individuals.

This generic methodology will be used in other food-related projects undertaken as part of the TRADINNOVATIONS project (Alzheimer and dysphagia patients) and other research projects carried out by the IATE Joint Research Unit, or even in scientific mediation. The collaboration with the Valentin Haüy association is expected to be renewed as part of students' projects and to help consolidate, improve, and validate the methodology. The Erasmus+ TRADINNOVATIONS project consortium will make it possible to test the methodology in six different European countries (France, Slovenia, Ireland, Spain, Portugal, Finland) as well as in Tunisia and Lebanon so as to integrate and compare cross-cultural specificities for the same studied target population. More broadly, the methodology used is part of a systemic study centered on the "eater" as an active actor in the conception of adapted food, linking it upstream to agricultural production and processing (on a gastronomic and industrial scale) and taking into account the constraints related to environment, climate change, and specific physiological and nutritional needs.

Final Conclusion and Future Outlook

The first part of this book proposes ontology-guided methods and tools designed to integrate heterogeneous data sources and meant to be coupled with decision support systems. The second part deals with methods and tools for collective decision support.

The domain ontologies put forward in Chapter 1 for food and biobased products provide a standardized vocabulary to incorporate the multiple dimensions of the agri-food system to support the design of sustainable, healthy, and affordable food and biobased products. The TransformON domain ontology is the only resource, to our knowledge, that covers both food and biobased products. Chapter 2 is dedicated to task ontologies for product processing and characterization processes, and more precisely details how the PO² ontology can be used to describe the different steps raw materials go through to become final products. It also captures the observations made on both the process steps and the materials/products themselves. Moreover, the CAPEX ontology represents expert knowledge and practical know-how. It helps suggest recommendations on what actions should be made to keep product quality high or to fix problems when defects appear during food or biobased product processing. The naRyQ ontology, an ontological and terminological resource (OTR), organizes data from heterogeneous external sources in the same way, so they can be compared and consistently linked. Finally, the @Web platform, which is built on naRyQ, also includes a model that assesses how reliable external data sources are before being reused in decision support systems. In Chapter 3, we presented an approach to use table and text mining techniques to semi-automatically feed a knowledge base with information extracted from a corpus of scientific articles. To this end, we rely on naRyQ and associated domain ontologies presented in Chapter 2 to guide the extraction. Two data structures, namely SciPuRe and STaRe, were designed to reconstitute instances of n-ary relations from text and tables. Chapter 4 addressed several decision-making problems by reusing data and knowledge bases presented in Chapter 2. Using collective know-how expressed as qualitative knowledge, the CAPEX platform proposes (by relying on the CAPEX ontology) technological recommendations to solve a production defect or achieve a given quality. The absence of a deterministic model can be managed from a reverse-engineering optimization perspective using machine learning models trained on data and knowledge available in the knowledge bases presented in Chapter 2. For decision-making purposes when using an incomplete data source containing candidate solutions also selected from knowledge bases presented in Chapter 2, it is possible to extract relevant information by using criteria considered as constraints or wishes thanks to fuzzy bipolar queries.

With regard to methods and tools for collective decision support, Chapter 5 covers methods that explicitly integrate arguments and justifications provided by stakeholders. Two complementary approaches are presented and compared: the DAMN tool, based on the statement graph formalism, and the MyChoice tool, which evaluates

arguments through their metadata. The first use case, ArgCW-LCA, illustrates how DAMN can be used to build transparent weighting profiles for life cycle assessment criteria, ensuring that stakeholder disagreements and justifications are taken into account before applying multicriteria analysis. MyChoice, in turn, is exemplified on the assessment of biowaste valorization scenarios, highlighting how visualization and analysis of argument metadata can reveal areas of consensus and conflict across diverse stakeholder groups. The chapter provides a comparative discussion of the two approaches, emphasizing their strengths, limitations, and complementarities. While DAMN prioritizes the content and logical structure of arguments, MyChoice facilitates broad participatory engagement and trend analysis. Together, they illustrate how argumentation-based methods can improve the inclusiveness, transparency, and legitimacy of decision-making processes in food chains. In Chapter 6, collective decision support is combined with scenario design to address whether innovations toward more sustainable systems would benefit society, which parts of it, in what manner, and under what conditions. This combination of scenario-building techniques with both ontologies and multicriteria/multistakeholder argumentation lays the groundwork for a new branch of social life cycle assessment (S-LCA) by considering the broader social, economic, and environmental consequences of innovations in the value chain, thus offering insights to assess future transitions. Shifting to the individual scale, Chapter 7 investigates how to evaluate the impact of an intervention on a population of participants with specific needs, through a generic methodology based on bipolar argumentation. It examines both expected and unexpected outcomes and includes short-term effects observed during the workshop and potential long-term impacts.

The main avenues for future research are also discussed in this book. Several knowledge bases were covered in Chapter 2, while Chapter 4 looked at various decision support models. Ongoing research is now focused on the interoperability between knowledge bases to build a federation that will bring new global added-value and services, such as allowing the enhancement of FAIR-by-Design implementation for experimental data management. Interoperability between knowledge bases that combine experimental data and technological expert know-how coupled with decision support models will also be developed in order to design digital twin models associated with bioresource processing.

Recent developments in AI and natural language processing have seen the rise of large language models (LLMs). Based on a transformer architecture, these models are trained on larger than ever text corpora in order to answer textual instructions (Toro et al., 2024). LLMs are clearly essential resources that will permit the enrichment of TransformON with concept textual definitions and facilitate the alignment of TransformON with other ontologies like FoodOn. LLMs will also help enhance the performance of table and text mining approaches presented in Chapter 3 to semi-automatically feed knowledge bases with information extracted from scientific articles.

Regarding Chapter 5, future developments could take several directions. For DAMN, one important challenge is to lower the barrier to entry for nonexpert users by hiding the use of formal logic behind natural language interfaces, which could be accomplished by using recent LLM technologies. It would also be worthwhile to build complete participatory platforms around DAMN, where functionalities are integrated into the tool to allow for live debate and the visualization of evolving consensus,

enabling broader stakeholder deliberation. For MyChoice, future work could focus on enriching the evaluation layer to check the consistency between arguments and aims, and extend its use in participatory forecasting, i.e., simulating how stakeholder attitudes and arguments might evolve over time or under new scenarios, thereby helping anticipate potential shifts in consensus. Together, these avenues point toward hybrid systems that combine rigorous reasoning with accessibility and inclusiveness. The methods in Chapters 6 and 7 can now be applied across other sectors and stabilized, thus contributing to more informed and collaborative decision-making processes for sustainable development. Ultimately, they pave the way for a more systematic approach to evaluating sustainability at all stakeholder levels, to gain a comprehensive understanding of the broader impacts of both innovations and interventions.

Appendix 1

Knowledge Engineering Models Used in the Book

Jérôme Fortin, Patrice Buche, Pierre Bisquert

» 1 Introduction

In the field of artificial intelligence, the ability to represent, structure, and reason about knowledge is essential for developing expert systems, intelligent agents, and decision support platforms. This appendix presents the description logics and Semantic Web standards that enable the formalization of knowledge in an interoperable and machine-readable way.

These models and associated tools are particularly useful in contexts where data is heterogeneous, distributed, and originates from diverse sources: technical documents, regulatory databases, observations, or expert input. For example, modeling the relationships between preservation methods, types of packaging, and microbiological risks requires a fine-grained structuring of concepts and their interactions.

Description logics address this need by relying on solid logical foundations, while offering languages adapted to publishing and exchanging knowledge on the Web.

In this appendix, we first introduce the logical foundations necessary to understand knowledge representation models, then present the main standards used in the Semantic Web (Antoniou and van Harmelen, 2004):

- RDF (Resource Description Framework), for structuring data as graphs.
- RDFS (RDF Schema), for enriching these graphs with vocabularies and hierarchies.
- OWL (Web Ontology Language), for modeling complex ontologies and enabling reasoning.
- SKOS (Simple Knowledge Organization System), for representing controlled vocabularies and thesauri.

Each section is illustrated with simple examples to show how these models and associated tools can be applied to concrete use cases such as traceability, expert knowledge structuring, or interoperability between systems.

» 2 Logical foundations

Formal knowledge representation relies on logical tools that allow information to be structured in a rigorous and machine-processable way. Before introducing Semantic

Web languages, it is useful to briefly recall the two main types of logic on which they are based: propositional logic and first-order logic.

2.1 Propositional logic

Propositional logic is the simplest logical system. It manipulates atomic propositions (e.g., P , Q , R) that can be true or false, and combines them using logical operators:

- $\neg P$ (not P)
- $P \wedge Q$ (P and Q)
- $P \vee Q$ (P or Q)
- $P \rightarrow Q$ (if P then Q)

This system is used to reason about simple facts but is limited: it cannot represent complex objects, relationships between objects, or properties.

Example:

Given the following propositions:

- P : “The product is perishable”
- Q : “The product is stored in a cold environment”
- R : “The product complies with storage requirements”

we could express a simple rule meaning that *if a product is perishable, then it must be stored cold; and if it is stored cold, then it complies with storage requirements*:

$$(P \rightarrow Q) \wedge (Q \rightarrow R)$$

2.2 First-order logic

First-order logic introduces quantifiers and relations between objects, called predicates. It allows the formulation of statements such as:

- $\forall x (\text{Fruit}(x) \rightarrow \text{Edible}(x))$

All fruits are edible

- $\exists x, \exists y (\text{Product}(x) \wedge \text{Plastic}(y), \text{PackagedIn}(x, y))$

There exists a product packaged in plastic.

This system enables the modeling of object classes, properties, binary or n-ary relations, and general rules. It is a powerful foundation for knowledge representation, but its expressiveness introduces important limitations for automated reasoning. First-order logic is semi-decidable: there exist statements for which a reasoning algorithm may never terminate, or may fail to prove all logical truths (issues of completeness). This means that a system based on this logic may be unable to answer certain queries, or to guarantee that all valid inferences will be produced.

2.3 Datalog and existential rules

First-order logic provides a very expressive framework for representing knowledge, but this expressiveness comes at the cost of quickly losing decidability. Indeed, reasoning in full first-order logic can lead to nonterminating processes, making it impractical for automated reasoning in knowledge-based systems. To address this, several restricted yet computationally controllable fragments of first-order logic have been proposed. Among them, Datalog and its existential rules extension play a central role.

Datalog is a rule-based language derived from first-order logic but without function symbols or existential quantifiers in rule heads. It is widely used in deductive databases and knowledge graph reasoning.

A Datalog rule has the general form of:

$$\forall X \forall Y B[X, Y] \rightarrow H[X]$$

where B and H are conjunctions of atoms, X is the set of variables shared by B and H , Y is the set of variables that occur only in B .

This means that if all the body atoms $B_i \in B$ hold, then the head atom H also holds. All variables appearing in the head must appear in the body, which ensures finite and safe inference. Existential rules extend this language by allowing existential quantifiers in rule heads, enabling the introduction of new, unnamed individuals when a rule is applied.

An existential rule has the form:

$$\forall X \forall Y B[X, Y] \rightarrow \exists Z. H[X, Z]$$

where B and H are conjunctions of atoms, X is the set of variables shared by B and H , Y is the set of variables that occur only in B and Z is the set of variables that occur only in H (and are, thus, existentially quantified).

For example:

$$\forall x \text{ person}(x) \rightarrow \exists y. \text{parent}(x, y)$$

states that every person has some parent, without specifying who that parent is.

The semantics of existential rules are defined through a constructive process known as the chase, which repeatedly applies rules to a database and introduces, if needed, fresh values for existential variables. Note that in general, this process may not terminate, so syntactic restrictions on existential rules (such as linear, guarded, or sticky, to name only a few) have been proposed to ensure decidability and tractable reasoning.

It is also interesting to note that existential rules generalize several description logic fragments (including DL-Lite and EL); the next section presents the notion of description logic.

2.4 Toward description logics

Description logics (DL) are a well-defined subset of first-order logic. They are designed to balance expressiveness and decidability, enabling efficient automated reasoning.

They constitute the formal foundation of languages such as OWL, and allow:

- the definition of concepts (classes), roles (properties), and individuals (basic entities of the domain that instantiate classes and can be connected to other individuals through roles)
- the expression of hierarchies, cardinality constraints, equivalences, etc.
- the execution of automated inferences (e.g., classification, inconsistency detection)

For example, one can express that a processed product is a product that has undergone at least one technological operation, or that a recyclable package is a package made of a recyclable material.

► 3 RDF – Resource Description Framework

The Resource Description Framework (RDF) is a data model standardized by the W3C⁴⁷ to represent structured and interoperable information on the Web. It is the foundational layer of the Semantic Web, enabling the description of resources and their relationships in the form of graphs.

The core principle of RDF is the triple: (subject, predicate, object). Each triple expresses a basic assertion: The subject is the resource being described. The predicate is the property or relation. The object is the value of that property (a literal or another resource).

Example 1:

(Product123, hasPackaging, Packaging456)

(Packaging456, isMadeOf, "plastic")

(Product123, hasName, "Plain yogurt")

These triples can be visualized as a directed graph, where nodes represent resources and edges represent relationships. In RDF, each resource is identified by a uniform resource identifier (URI), ensuring uniqueness and facilitating data interlinking at Web scale.

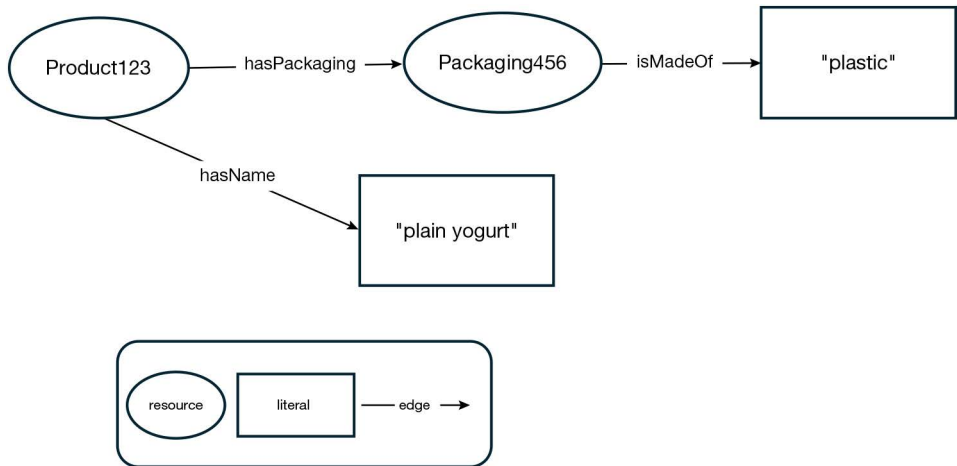


Figure A1.1. Graphical representation of the RDF graph associated with Example 1.

The object of a triple can be a resource (identified by a URI) or a literal (text, number, date, etc.).

Example 2:

In Figure A1.1, Product123 and Packaging456 are two resources, while "plastic" and "plain yogurt" are textual literal values.

RDF can be expressed in several syntaxes: Turtle (human-readable), RDF/XML, JSON-LD, N-Triples, etc. Although RDF does not support complex reasoning on

47. RDF (2014, February 25; <https://www.w3.org/RDF/>).

its own, it provides a normalized data structure upon which more expressive layers (such as RDFS and OWL) can be built. It also supports SPARQL, the standard query language for RDF graphs.⁴⁸

►► 4 RDFS – RDF Schema

The RDF model allows facts to be represented as triples, but it does not provide mechanisms to semantically structure or organize these facts. This is the role of RDFS (RDF Schema), which extends RDF by introducing vocabulary, typing, and hierarchy mechanisms.⁴⁹ RDFS allows the definition of classes (types), properties, and relationships between these elements, making it a simple but limited semantic structuring language.

In RDFS, two main kinds of relations (or properties) can be distinguished. First are structural relations, such as `subClassOf` and `subPropertyOf`, which define hierarchies between classes and between properties to allow for the construction of taxonomies and inheritance of characteristics. The second type of relations are semantic relations used as predicates in RDF triples (e.g., *hasPackaging*), which specify the meaning of directed links between a subject and an object. Such predicates connect instances of classes – for example, a *Product* and a *Packaging* – and thereby provide the semantic content of the graph.

For example, one can declare that:

- Product is a class
- Packaging is a class
- hasPackaging is a property

This property links a Product to a Packaging.

RDFS also allows the construction of class and property hierarchies through the property *rdfs:subClassOf*. For example, this means that every *DairyProduct* is a *Product*, and that *hasFoodPackaging* is a specialization of *hasPackaging*. Thanks to these mechanisms, RDFS enables the structuring of a vocabulary and the inference of implicit facts. For instance, if an individual is declared as a *DairyProduct*, an RDF engine can infer that it is also a *Product*. However, RDFS remains limited in expressiveness: it cannot express complex constraints (such as cardinalities, disjunctions, or class equivalences). For that, more powerful languages like OWL are used, which we will explore in the next section.

►► 5 OWL – Web Ontology Language

The Web Ontology Language (OWL) is a W3C standard designed to model rich and formally defined ontologies.⁵⁰ It is based on description logics, which provide formal semantics and enable automated reasoning over the represented knowledge. Unlike RDF and RDFS, which allow the structuring of data graphs, OWL supports the expression of logical constraints, complex definitions, and inference rules.

OWL introduces several key notions:

- Classes: sets of individuals (e.g., *Product*, *Packaging*)
- Individuals: concrete instances (e.g., *Yogurt123*, *Bottle456*)

48. SPARQL 1.1 Query Language (2013, March 21; <https://www.w3.org/TR/sparql11-query/>).

49. RDF Schema 1.1 (2014, February 25; <https://www.w3.org/TR/rdf-schema/>).

50. OWL (2012, December 11; <https://www.w3.org/OWL/>).

- Properties: relationships between individuals or between an individual and a value (e.g., `hasPackaging`, `isMadeOf`)
- Axioms: logical statements about classes, properties, or individuals

OWL requires that properties are either `ObjectProperties` or `DatatypeProperties`: `DatatypeProperties` are relations between instances of classes and literals, while `ObjectProperties` are relations between instances of two classes.

OWL allows also the definition of class equivalences, restrictions (e.g., cardinality, value types), defined classes with necessary and sufficient conditions, and property characteristics (e.g., transitive, symmetric, functional).

Let's consider a simple example:

```
ex:Product a owl:Class .
ex:Packaging a owl:Class .
ex:hasPackaging a owl:ObjectProperty ;
    rdfs:domain ex:Product ;
    rdfs:range ex:Packaging .
```

We can define a class by restriction that states that a *ProductWithPlasticPackaging* is a product that has at least one packaging made of plastic:

```
ex:ProductWithPlasticPackaging a owl:Class ;
    owl:equivalentClass [
        a owl:Class ;
        owl:intersectionOf (
            ex:Product
            [ a owl:Restriction ;
                owl:onProperty ex:hasPackaging ;
                owl:someValuesFrom [
                    a owl:Class ;
                    owl:hasValue "plastic" ;
                    owl:onProperty ex:isMadeOf
                ]
            ]
        )
    ] .
```

Thanks to OWL's formal semantics, a reasoner can infer class membership for individuals, check ontology consistency, and deduce implicit relationships (e.g., hierarchy, equivalence).

For example, if we declare:

```
ex:Yogurt123 a ex:Product .
ex:Yogurt123 ex:hasPackaging ex:Bottle456 .
ex:Bottle456 ex:isMadeOf "plastic" .
```

a reasoner can automatically infer that *Yogurt123* is an instance of *ProductWithPlasticPackaging*.

OWL is available in several profiles, offering different trade-offs between expressiveness and computational performance:

- OWL Lite: a simplified version, rarely used today

- OWL DL: based on Description Logics, with complete and decidable reasoning
- OWL Full: highly expressive, but without guarantees of decidability

In practice, OWL DL (and its new version, OWL 2) is the most commonly used in systems requiring reliable reasoning.

► 6 SKOS – Simple Knowledge Organization System

SKOS is an RDF-based model designed to represent knowledge organization systems such as thesauri, taxonomies, subject headings, or controlled vocabularies. It is standardized by the W3C and fits within the Semantic Web framework, while remaining simpler and more flexible than OWL.⁵¹

Unlike OWL, which supports complex logical modeling, SKOS focuses on structuring vocabularies in a way that is interoperable and human-readable. The central element in SKOS is the concept (*skos:Concept*), representing a unit of meaning (e.g., Dairy Product, Recyclable Packaging, Storage Temperature). Each concept can be described using:

- a preferred label (*skos:prefLabel*): the main name of the concept
- alternative labels (*skos:altLabel*): synonyms or lexical variants
- a definition (*skos:definition*): textual description
- hierarchical relations (*skos:broader*, *skos:narrower*): generalization/specialization
- associative relations (*skos:related*): non-hierarchical links

Example:

ex:Yogurt123 rdf:type skos:Concept .

ex:Yogurt123 skos:prefLabel "yogurt"@en .

ex:Yogurt123 skos:altLabel "yoghurt"@en .

ex:Yogurt123 skos:prefLabel "yaourt"@fr .

ex:Yogurt123 skos:altLabel "yoghourt"@fr .

ex:Yogurt123 skos:definition "a creamy, slightly sour food made by fermenting milk with specific bacteria, usually eaten plain or flavored, and often considered a source of probiotics." @en .

ex:Yogurt123 skos:definition "aliment lacté obtenu par fermentation du lait grâce à des bactéries spécifiques, de texture crémeuse et de goût légèrement acidulé, consommé nature ou aromatisé." @fr .

SKOS is particularly suited for publishing or sharing domain-specific vocabularies, structuring data categories in information systems, aligning multiple vocabularies (via *skos:exactMatch*, *skos:closeMatch*, etc.), or enabling semantic navigation in data portals or digital libraries. It is widely used in digital libraries, archives, museums, open data platforms, and scientific repositories (e.g., AGROVOC, EuroVoc, MeSH). SKOS does not support complex logical reasoning. It does not define classes, properties, or formal constraints. It is therefore complementary to OWL: SKOS can be used to structure a vocabulary, while OWL can model logical relationships between concepts.

51. Simple Knowledge Organization System Reference (2009, August 18; <https://www.w3.org/TR/skos-reference/>).

» 7 Fuzzy sets

Fuzzy sets, introduced by Zadeh in 1965 (Zadeh, 1965), provide a mathematical framework for representing imprecision, preferences, and similarities by way of gradual membership. Unlike classical sets, where an element either does or does not belong to a set, fuzzy sets allow intermediate degrees of membership between 0 and 1. This makes them well suited for modeling vague concepts or expressing user preferences in information systems.

A fuzzy set μ defined on a space A is a function $\mu: A \rightarrow [0,1]$ where $\mu(x)$ represents the membership degree of element x .

The support $S(\mu)$ is the set of $x \in A$ for which $\mu(x) > 0$, and the kernel $K(\mu)$ is the set of x for which $\mu(x) = 1$.

Examples from @Web

In the @Web data warehouse, fuzzy sets on discrete domains are used to enrich queries expressing preferences as filtering criteria. Consider the discrete domain of microorganism types {Gram+, Gram-}. A user can express preferences as a fuzzy set:

$$\mu_{MicroPref} = \{(Gram+, 1.0), (Gram-, 0.5)\}.$$

where $support(\mu_{MicroPref}) = \{(Gram+, 1.0)\}$ and $kernel(\mu_{MicroPref}) = \{\}$.

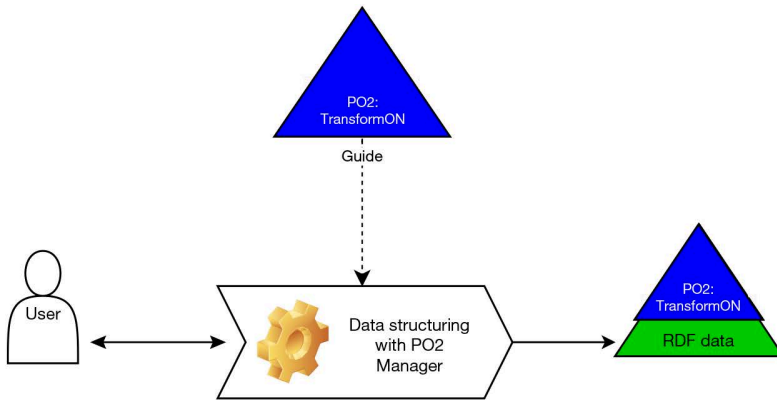
Thus, in a query aimed at retrieving growth parameters, a result concerning a Gram+ microorganism will be considered fully satisfactory ($\mu=1$), whereas a result concerning a Gram- microorganism will still be accepted, but with reduced satisfaction ($\mu=0.5$)

Appendix 2

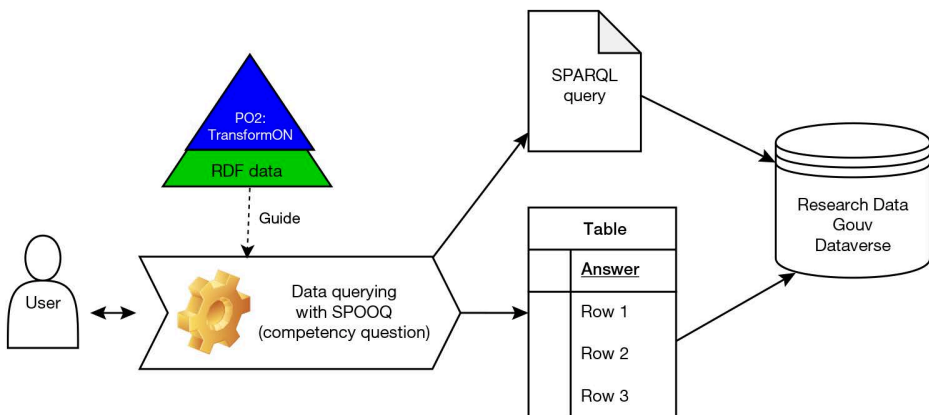
Legend for the Workflow Diagrams

Patrice Buche

Workflow diagrams are used in all the chapters to describe the software workflows used to implement the proposed solutions to the research questions. The legend used in those diagrams (based on the example of Figure 2.2 replicated below) is presented here.



(a) Workflow to enrich the RDF data part of PO2 knowledge base



(b) Workflow to query the RDF data part of the PO2 knowledge base

Figure A2.1. Workflows associated with the PO² knowledge base.

In this figure, the blue triangles represent ontologies. In Figure A2.1, they represent the PO² task ontology presented in Chapter 2 specialized by the TransformON domain ontology presented in Chapter 1.

The green parts of the triangles represent RDF graphs that correspond to instances of classes and properties belonging to the ontologies. In Figure A2.1, they correspond to the RDF graph instances of the TransformON ontology.

The rectangular arrows with a cog represent semi-automatic processing. In Figure A2.1, the first rectangular arrow (a) represents the semi-automatic processing performed by the PO² Manager software to generate RDF graphs, instances of the TransformON ontology. The second rectangular arrow (b) represents the data querying step performed using SPO2Q software, which generates a SPARQL query and the data table in CSV format corresponding to the query execution on the PO² triple store.

Cylinders represent databases and data warehouses excluding RDF databases represented as parts of triangles as already explained above. In Figure A2.1, the cylinder represents the Research Data Gouv dataverse.

References

- Abou Jaoudé L., Charrier J., Darcel N.N., Maurice A., Perrin L., Thomopoulos R., 2022. Un état des lieux des messages alimentaires à destination des enfants [Conference poster]. Journées Francophones de Nutrition, Toulouse, France. <https://hal.inrae.fr/hal-03856733>
- Aceves Lara C.A., Athès V., Buche P., Della Valle G., Farines V., Fonseca F. et al., 2018. The virtual food system: Innovative models and experiential feedback in technologies for winemaking, the cereals chain, food packaging and eco-designed starter production. *Innovative Food Science & Emerging Technologies*, 46, 54–64. <https://doi.org/10.1016/j.ifset.2017.10.006>
- Aleksovski Z., Klein M., ten Kate W., van Harmelen F., 2006. Matching unstructured vocabularies using a background ontology. In: S. Staab, V. Svátek (Eds.), *Managing knowledge in a world of networks – EKAW 2006. Lecture notes in computer science*, 4248, 182–197, Springer Berlin Heidelberg. https://doi.org/10.1007/11891451_18
- Ali S., Abuhmed T., El-Sappagh S., Muhammad K., Alonso-Moral J.M., Confalonieri R. et al., 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- Amgoud L., Bonnefon J.-F., Prade H., 2005. An argumentation-based approach to multiple criteria decision. In: L. Godo (Ed.), *Symbolic and quantitative approaches to reasoning with uncertainty. ECSQARU 2025. Lecture notes in computer science*, 3571, 269–280. https://doi.org/10.1007/11518655_24
- Amgoud L., Prade H., 2009. Using arguments for making and explaining decisions. *Artificial Intelligence*, 173(3–4), 413–436. <https://doi.org/10.1016/j.artint.2008.11.006>
- Annane A., Bellahsene Z., Azouaou F., Jonquet C., 2018. Building an effective and efficient background knowledge resource to enhance ontology matching. *Journal of Web Semantics*, 51, 51–68. <https://doi.org/10.1016/j.websem.2018.04.001>
- Antoniou G., van Harmelen F., 2004. *A Semantic Web Primer*. MIT Press, 272 p.
- Arp R., Smith B., Spear A.D., 2015. *Building Ontologies with Basic Formal Ontology*. MIT Press, 248 p.
- Aubin S., Bernard É., Bravo S., Cadiou C., Girard A., Weber M., 2024. Révéler et valoriser vos données avec le thésaurus INRAE. *NOVAE*, 3. <https://doi.org/10.17180/NOVAE-2024-NO-ART03>
- Baig A.A., Ruzli R., Buang A.B., 2013. Reliability analysis using fault tree analysis: a review. *International Journal of Chemical Engineering and Applications*, 4(3), 169–173. <https://doi.org/10.7763/IJCEA.2013.V4.287>
- Bandrowski A., Brinkman R., Brochhausen M., Brush M.H., Bug B., Chibucos M.C. et al., 2016. The Ontology for Biomedical Investigations. *PLOS ONE*, 11(4), e0154556. <https://doi.org/10.1371/journal.pone.0154556>
- Bandura A., 1997. *Self-Efficacy: The Exercise of Control*. W.H. Freeman and Company, 604 p.
- Barbar R., Malbec J.-M., Lavelle C., Kientza H.T. vo. 2021. An educational satellite project around the scientific elucidation of culinary precisions in Lebanon and in the Middle East. In: C. Lavelle, H. This, A.L. Kelly, R. Burke (Eds.), *Handbook of molecular gastronomy – Scientific foundations, educational practices, and culinary applications*. CRC Press, 894 p. <https://doi.org/10.1201/9780429168703>
- Barré R., 2000. Le foresight britannique : Un nouvel instrument de gouvernance ? *Futuribles*, 249, 5–24.
- Basco-Carrera L., Warren A., van Beek E., Jonoski A., Giardino A., 2017. Collaborative modelling or participatory modelling? A framework for water resources management. *Environmental Modelling & Software*, 91, 95–110. <https://doi.org/10.1016/j.envsoft.2017.01.014>
- Baudrit C., Perrot N., Brousset J.M., Abbal P., Guillemain H., Perret B. et al., 2015. A probabilistic graphical model for describing the grape berry maturity. *Computers and Electronics in Agriculture*, 118, 124–135. <https://doi.org/10.1016/j.compag.2015.08.019>

- Baudrit C., Buche P., Leconte N., Fernandez C., Belna M., Gésan-Guiziou G., 2022. Decision support tool for the agri-food sector using data annotated by ontology and Bayesian network: a proof of concept applied to milk microfiltration. *International Journal of Agricultural and Environmental Information Systems*, 13(1), 1–22. <https://doi.org/10.4018/IJAEIS.309136>
- Baudrit C., Fernandez C., Couteaux J., Ndiaye A., 2024. Electronic knowledge books (eK-Books) as a medium to capitalise on and transfer scientific, engineering, operational, technological and craft knowledge. *PLOS ONE*, 19(5), e0299150. <https://doi.org/10.1371/journal.pone.0299150>
- Baudrit C., Fernandez C., Couteaux J., Buche P., Bel N., Charles C., Notz E., 2025. Cheese eK-Book: a new web-based medium for capitalising on, structuring and transferring cheesemaking knowledge and know-how. *Journal of Documentation*, 81(2), 369–384. <https://doi.org/10.1108/JD-07-2024-0182>
- Belaud J.-P., Prioux N., Vialle C., Buche P., Destercke S., Barakat A., Sablayrolles C., 2022. Intensive data and knowledge-driven approach for sustainability analysis: application to lignocellulosic waste valorization processes. *Waste and Biomass Valorization*, 13, 583–598. <https://doi.org/10.1007/s12649-021-01509-8>
- Belna M., Ndiaye A., Taillandier F., Fernandez C., Agabriel L., Gésan-Guiziou G., 2022. Multiobjective optimization of skim milk microfiltration based on expert knowledge. *Expert Systems with Applications*, 205, 117624. <https://doi.org/10.1016/j.eswa.2022.117624>
- Belton V., Stewart T., 2010. Problem structuring and multiple criteria decision analysis. In: M. Ehrgott, J.R. Figueira, S. Greco (Eds.), *Trends in Multiple Criteria Decision Analysis. International Series in Operations Research & Management Science*, Springer, 209–239. https://doi.org/10.1007/978-1-4419-5904-1_8
- Bench-Capon T.J., 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3), 429–448. <https://doi.org/10.1093/logcom/13.3.429>
- Berrahou S.L., Buche P., Dibie-Barthelemy J., Roche M., 2015. Identification des unités de mesure dans les textes scientifiques. In: J.-M. Lecarpentier, N. Lucas (Eds.), *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*. [Conference proceedings], 88–94. <https://aclanthology.org/2015.jeptalnrecital-court.14/>
- Berrahou S.L., Buche P., Dibie J., Roche M., 2017. Xart: discovery of correlated arguments of n-ary relations in text. *Expert Systems with Applications*, 73, 115–124. <https://doi.org/10.1016/j.eswa.2016.12.028>
- Berthet M.-A., Angellier-Coussy H., Chea V., Guillard V., Gastaldi E., Gontard N., 2015. Sustainable food packaging: valorising wheat straw fibres for tuning PHBV-based composites properties. *Composites Part A: Applied Science and Manufacturing*, 72, 139–147. <https://doi.org/10.1016/j.compositesa.2015.02.006>
- Besnard P., Hunter A., 2001. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1–2), 203–235. [https://doi.org/10.1016/S0004-3702\(01\)00071-6](https://doi.org/10.1016/S0004-3702(01)00071-6)
- Bhagavatula C.S., Noraset T., Downey D., 2015. TabEL: entity linking in web tables. In: M. Arenas, O. Corcho, E. Simperl et al. (Eds.), *The Semantic Web – ISWC 2015. Lecture Notes in Computer Science*, 9366, 425–441, Springer. https://doi.org/10.1007/978-3-319-25007-6_25
- Bilyk M.C., Sontrop J.M., Chapman G.E., Barr S.I., Mamer L., 2009. Food experiences and eating patterns of visually impaired and blind people. *Canadian Journal of Dietetic Practice and Research*, 70(1), 13–18. <https://doi.org/10.3148/70.1.2009.13>
- Bisquert P., Buche P., Hecham A., Koptelov M., 2024. Justified Preference Aggregation in Agri-Food Systems: An approach, Argumentation Methods and Tools (Supplementary Material). <https://hal.science/hal-04607020v1>
- Bisquert P., Hecham A., Buche P., Koptelov M., 2025. Justified preference aggregation in agri-food systems: an approach, argumentation methods, and tools. *International Journal of Agricultural and Environmental Information Systems*, 16(1), 1–26. <https://doi.org/10.4018/IJAEIS.393045>
- Bosc P., Lietard L., Pivert O., 1994. Soft querying, a new feature for database management systems. In: D. Karagiannis (Ed.), *Database and Expert Systems Applications. Lecture Notes in Computer Science*, 856, 631–640, Springer. https://doi.org/10.1007/3-540-58435-8_229
- Bosc P., Pivert O., 1995. SQLf: a relational database language for fuzzy querying. *IEEE Transactions on Fuzzy Systems*, 3(1), 1–17. <https://doi.org/10.1109/91.366566>

- Bourguet J.-R., Thomopoulos R., Mugnier M.-L., Abécassis J., 2013. An artificial intelligence-based approach to deal with argumentation applied to food quality in a public health policy. *Expert Systems with Applications*, 40(11), 4539–4546. <https://doi.org/10.1016/j.eswa.2013.01.059>
- Bourigault D., Jacquemin C., 1999. Term extraction + term clustering: an integrated platform for computer-aided terminology. *EACL '99: Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 15–22. <https://doi.org/10.3115/977035.977039>
- Boyce B.R., Meadow C.T., Kraft D.H., Barry C., 2007. *Text information retrieval systems*. Elsevier, 371 p.
- Brinkman R.R., Courtot M., Derom D., Fostel J.M., He Y., Lord P., et al., the OBI consortium, 2010. Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*, 1(S1), S7. <https://doi.org/10.1186/2041-1480-1-S1-S7>
- Bruno N., Chaudhuri S., Gravano L., 2002. Top-k selection queries over relational databases: mapping strategies and performance evaluation. *ACM Transactions on Database Systems*, 27(2), 153–187. <https://doi.org/10.1145/568518.568519>
- Buche P., Dibie-Barthelemy J., Ibanescu L., Soler L., 2013a. Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 805–819. <https://doi.org/10.1109/TKDE.2011.245>
- Buche P., Dervaux S., Dibie-Barthelemy J., Ibanescu L., Soler L., Touhami R., 2013b. Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique. *Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle*, 27(4–5), 539–568. <https://hal-lirmm.ccsd.cnrs.fr/lirmm-00903768v1>
- Buche P., Cuq B., Fortin J., Sipieter C., 2019. Expertise-based decision support for managing food quality in agri-food companies. *Computers and Electronics in Agriculture*, 163, 104843. <https://doi.org/10.1016/j.compag.2019.05.052>
- Buche P., Cufi J., Dervaux S., Dibie J., Ibanescu L., Oudot A., Weber M., 2021a. How to manage incompleteness of nutritional food sources? A solution using FoodOn as pivot ontology. *International Journal of Agricultural and Environmental Information Systems*, 12(4), 1–26. <https://doi.org/10.4018/IJAEIS.20211001.0a4>
- Buche P., Dervaux S., Leconte N., Belna M., Granger-Delacroix M., Garnier-Lambrouin F. et al., 2021b. Milk microfiltration process dataset annotated from a collection of scientific papers. *Data in Brief*, 36, 107063. <https://doi.org/10.1016/j.dib.2021.107063>
- Buche P., Couteaux J., Cufi J., Destercke S., Oudot A., 2022. Food quality decision tree based on collective know-how (Capex ontology) [Dataset]. Recherche Data Gouv, V4. <https://doi.org/10.57745/SEJP1B>
- Buche P., Couteaux J., Cufi J., Destercke S., Oudot A., 2023. Integrating collective know-how for multicriteria decision support in agrifood chains—application to cheesemaking. *Frontiers in Artificial Intelligence*, 6, 1145007. <https://doi.org/10.3389/frai.2023.1145007>
- Cacioppo J.T., Gardner W.L., Berntson G.G., 1997. Beyond bipolar conceptualizations and measures: the case of attitudes and evaluative space. *Personality and Social Psychology Review*, 1(1), 3–25. https://doi.org/10.1207/s15327957pspr0101_2
- Cafarella M.J., Downey D., Soderland S., Etzioni O., 2005. Knowitnow: Fast, scalable information extraction from the web. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 563–570.
- Cagnon T., Guillaume C., Guillard V., Gontard N., 2013. Nanostructuring and microstructuring of materials from a single agropolymer for sustainable MAP Preservation of Fresh Food. *Packaging Technology and Science*, 26(3), 137–148. <https://doi.org/10.1002/pts.1973>
- Cannarella C., Piccioni V., 2011. Traditionations: Creating innovation from the past and antique techniques for rural areas. *Technovation*, 31(12), 689–699. <https://doi.org/10.1016/j.technovation.2011.07.005>
- Cayrol C., Lagasque-Schiex M.C., 2005. On the acceptability of arguments in bipolar argumentation frameworks. In: L. Godo (Ed.), *Symbolic and Quantitative Approaches to Reasoning with Uncertainty. ECSQARU 2005. Lecture Notes in Computer Science*, 3571, 378–389, Springer. https://doi.org/10.1007/11518655_33

- Chaib R.L., Macombe C., Thomopoulos R., 2021. Adaptation of a participatory system-modeling method to the constraints of remote working. Conference on Complex Systems – France, Dijon. <https://hal.science/hal-03770645>
- Chaib R.L., Macombe C., Thomopoulos R., 2022. Structuring ontologies from natural language for collaborative scenario modeling in agri-food systems. *Frontiers in Artificial Intelligence*, 5, 1056989. <https://doi.org/10.3389/frai.2022.1056989>
- Chaib R.L., 2023. Méthodologie pour anticiper et évaluer les impacts en avance : combiner l'ACV sociale et l'argumentation multicritère [PhD thesis, INRAE]. <https://theses.hal.science/tel-04448150>
- Chaib R.L., Macombe C., Thomopoulos R., 2023a. Adapting a participatory modelling method to forecast food system scenarios: a case study on the pork value-chain. *Economia agro-alimentare*, 24(3), 1–37. <https://doi.org/10.3280/ecag2022oa14488>
- Chaib R.L., Macombe C., Thomopoulos R., 2023b. Ontology: impacts of changes in value chains [Dataset]. Recherche Data Gouv, V1. <https://doi.org/10.57745/YHSDWA>
- Chaib R.L., Macombe C., Thomopoulos R., 2023c. Scénarios possibles d'évolution de la filière porcine française. *La revue française de la recherche en viandes et produits carnés*, <https://hal.inrae.fr/hal-04206445v1>
- Chaib R.L., Thomopoulos R., Macombe C., 2023d. Structuring ontologies in a context of collaborative system modelling. *arXiv*. <https://doi.org/10.48550/arXiv.2301.05478>
- Charles F., Sanchez J., Gontard N., 2003. Active modified atmosphere packaging of fresh fruits and vegetables: modeling with tomatoes and oxygen absorber. *Journal of Food Science*, 68(5), 1736–1742. <https://doi.org/10.1111/j.1365-2621.2003.tb12321.x>
- Charles F., Anchez J.S., Gontard N., 2005. Modeling of active modified atmosphere packaging of endives exposed to several postharvest temperatures. *Journal of Food Science*, 70(8), e443–e449. <https://doi.org/10.1111/j.1365-2621.2005.tb11512.x>
- Chomicki J., 2003. Preference formulas in relational queries. *ACM Transactions on Database Systems*, 28(4), 427–466. <https://doi.org/10.1145/958942.958946>
- Church K.W., Hanks P., 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.
- Cohen W.W., Ravikumar P., Fienberg S.E. et al., 2003. A comparison of string distance metrics for name-matching tasks. *IIWEB'03: Proceedings of the 2003 International Conference on Information Integration on the Web*, 3, 73–78.
- Cohen K.B., Johnson H.L., Verspoor K., Roeder C., Hunter L.E., 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11, 492. <https://doi.org/10.1186/1471-2105-11-492>
- Cooper W.S., 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1), 19–37. [https://doi.org/10.1016/0020-0271\(71\)90024-6](https://doi.org/10.1016/0020-0271(71)90024-6)
- Corrales D.C., Corrales J.C., Ledezma A., 2018a. How to address the data quality issues in regression models: a guided process for data cleaning. *Symmetry*, 10(4), 99. <https://doi.org/10.3390/sym10040099>
- Corrales D.C., Ledezma A., Corrales J.C., 2018b. From theory to practice: a data quality framework for classification tasks. *Symmetry*, 10(7), 248. <https://doi.org/10.3390/sym10070248>
- Corrales D.C., Ledezma A., Corrales J.C., 2020. A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks. *Applied Soft Computing*, 90, 106180. <https://doi.org/10.1016/j.asoc.2020.106180>
- Craswell N., 2009. Precision at n. In: L. Liu, M.T. Özsu (Eds.), *Encyclopedia of Database Systems*, 2127–2128, Springer. https://doi.org/10.1007/978-0-387-39940-9_484
- David G., Vannini M., Sisti L., Marchese P., Celli A., Gontard N., Angellier-Coussy H., 2020. Eco-conversion of two winery lignocellulosic wastes into fillers for biocomposites: vine shoots and wine pomaces. *Polymers*, 12(7), 1530. <https://doi.org/10.3390/polym12071530>
- David G., Croxatto Vega G., Sohn J., Nilsson A.E., Hélias A., Gontard N., Angellier-Coussy H., 2021. Using life cycle assessment to quantify the environmental benefit of upcycling vine shoots as fillers in biocomposite packaging materials. *The International Journal of Life Cycle Assessment*, 26(4), 738–752. <https://doi.org/10.1007/s11367-020-01824-7>

- Defossez J., Buche P., Guinet N., Guyonvarch Y., Todorov K., 2022. Oqlassif: Catégorisation semi-automatique de l'offre alimentaire en familles Oqali. INRAE, 23 p. <https://hal.inrae.fr/hal-05507173>
- De Jong E., Higson A., Walsh P., Wellisch M., 2012. Bio-based chemicals: Value added products from biorefineries – IEA Bioenergy – Task 42 Biorefinery. <https://www.ieabioenergy.com/blog/publications/bio-based-chemicals-value-added-products-from-biorefineries/>
- De Tre G., Zadrozny S., Bronselaer A.J., 2010. Handling bipolarity in elementary queries to possibilistic databases. *IEEE Transactions on Fuzzy Systems*, 18(3), 599–612. <https://doi.org/10.1109/TFUZZ.2010.2044183>
- Delhomme B., 2022. A web application based on argumentation frameworks and MCDA methods for real-time decision making – application to buildings retrofitting [Doctoral dissertation, Université de Bordeaux]. <https://theses.hal.science/tel-03635399v1>
- Delhomme B., Taillandier F., Abi-Zeid I., Thomopoulos R., Baudrit C., Mora L., 2022. An interface between natural language and abstract argumentation frameworks for real-time debate analysis. *Decision Support Systems*, 154, 113694. <https://doi.org/10.1016/j.dss.2021.113694>
- Destercke S., Buche P., Guillard V., 2011. A flexible bipolar querying approach with imprecise data and guaranteed results. *Fuzzy Sets and Systems*, 169(1), 51–64. <https://doi.org/10.1016/j.fss.2010.12.014>
- Destercke S., Buche P., Charnomordic B., 2013. Evaluating data reliability: an evidential answer with application to a web-enabled data warehouse. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 92–105. <https://doi.org/10.1109/TKDE.2011.179>
- Devlin J., Chang M.-W., Lee K., Toutanova K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. <https://arxiv.org/abs/1810.04805>
- Dice L.R., 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>
- Dooley D.M., Griffiths E.J., Gosal G.S. et al., 2018. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2, 23. <https://doi.org/10.1038/s41538-018-0032-6>
- Dooley D., Naravane T., 2023. Ontological how and why: action and objective of planned processes in the food domain. *Frontiers in Artificial Intelligence*, 6, 1137961. <https://doi.org/10.3389/frai.2023.1137961>
- Dooley D., Weber M., Ibanescu L., Lange M., Chan L., Soldatova L. et al., 2024. Food process ontology requirements. *Semantic Web – Interoperability, Usability, Applicability*, 15(4), 1133–1164. <https://doi.org/10.3233/SW-223096>
- Dubois D., Prade H., 1988. Possibility theory: an approach to computerized processing of uncertainty. Plenum Press, 287 p.
- Dubois D., Prade H., 1997. The three semantics of fuzzy sets. *Fuzzy Sets and Systems*, 90(2), 141–150. [https://doi.org/10.1016/S0165-0114\(97\)00080-8](https://doi.org/10.1016/S0165-0114(97)00080-8)
- Dubois D., Lorini E., Prade H., 2013. Bipolar possibility theory as a basis for a logic of desires and beliefs. In: W. Liu, V.S. Subrahmanian, J. Wijsen (Eds.), Scalable Uncertainty Management. *SUM 2013. Lecture Notes in Computer Science*, 8078, 204–218, Springer. https://doi.org/10.1007/978-3-642-40381-1_16
- Dubois D., Prade H., 2009. An overview of the asymmetric bipolar representation of positive and negative information in possibility theory. *Fuzzy Sets and Systems*, 160(10), 1355–1366. <https://doi.org/10.1016/j.fss.2008.11.006>
- Dung P.M., 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2), 321–357. [https://doi.org/10.1016/0004-3702\(94\)00041-X](https://doi.org/10.1016/0004-3702(94)00041-X)
- Duong T.H., Nguyen H.Q., Jo G.S., 2017. Smart data: where the big data meets the semantics. *Computational Intelligence and Neuroscience*, 2017, 6925138. <https://doi.org/10.1155/2017/6925138>
- Eckart de Castilho R., Mújdricza-Maydt É., Yimam S.M., Hartmann S., Gurevych I., Frank A., Biemann C., 2016. A web-based tool for the integrated annotation of semantic and syntactic structures. In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 76–84. <https://www.aclweb.org/anthology/W16-4011>

- EFSA (European Food Safety Authority), 2013. Standard Sample Description ver. 2.0. *EFSA Journal*, 11(10). <https://doi.org/10.2903/j.efsa.2013.3424>
- EFSA (European Food Safety Authority), 2015. The food classification and description system FoodEx 2 (revision 2). *EFSA Supporting Publications*, 12(5). <https://doi.org/10.2903/sp.efsa.2015.EN-804>
- Eftimov T., Korošec P., Koroušić Seljak B., 2017. StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients*, 9(6), 542. <https://doi.org/10.3390/nu9060542>
- Ehrmann M., Della Rocca L., Steinberger R., Tannev H., 2013. Acronym recognition and processing in 22 languages. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, 237–244. <https://aclanthology.org/R13-1031>
- Engel E., Planche C., Thomopoulos R., Macombe C., Nielsen Leroux C., Bazoché P. et al., 2022. INSECT4CITY: benefits and risks of insect-based bioconversion for recycling bio-waste from urban and peri-urban areas. Insects to Feed the World (IFW 2022). Quebec, Canada. <https://hal.science/hal-03696802>
- Euzenat J., Mochol M., Shvaiko P., Stuckenschmidt H., Svab O., Svátek V. et al., 2006. Results of the ontology alignment evaluation initiative 2006. In: *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006 and ISWC-2006)*, 225, 73–95.
- Fabre C., Buche P., Rouau X., Mayer-Laigle C., 2020. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in Brief*, 33, 106430. <https://doi.org/10.1016/j.dib.2020.106430>
- Faria D., Pesquita C., Santos E., Palmonari M., Cruz I.F., Couto F.M., 2013. The AgreementMakerLight Ontology Matching System. In: R. Meersman, H. Panetto, T. Dillon et al. (Eds.), *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, 8185, 527–541, Springer. https://doi.org/10.1007/978-3-642-41030-7_38
- Fishbein M., 1967. A behavior theory approach to the relations between beliefs about an object and the attitude toward the object. In: M. Fishbein (Ed.), *Readings in attitude theory and measurement*, 389–400, Wiley, New York. <https://amsr.contentdm.oclc.org/digital/collection/p21050coll5/id/4962/>
- Foppiano L., Romary L., Ishii M., Tanifuji M., 2019. Automatic identification and normalisation of physical measurements in scientific literature. *DocEng'19: Proceedings of the ACM Symposium on Document Engineering 2019*, 1–4. <https://doi.org/10.1145/3342558.334541>
- Franco-Salvador M., Rosso P., Navigli R., 2014. A knowledge-based representation for cross-language document retrieval and categorization. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 414–423. <https://doi.org/10.3115/v1/E14-1044>
- Freire P., Freire D., Licon C.C., 2025. A comprehensive review of machine learning and its application to dairy products. *Critical Reviews in Food Science and Nutrition*, 65(10), 1878–1893. <https://doi.org/10.1080/10408398.2024.2312537>
- Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y. et al., 2024. Retrieval-augmented generation for large language models: a survey. *arXiv*. <https://doi.org/10.48550/arXiv.2312.10997>
- García A.J., Simari G.R., 2004. Defeasible logic programming: an argumentative approach. *Theory and Practice of Logic Programming*, 4(1–2), 95–138. <https://doi.org/10.1017/S1471068403001674>
- Ghrab S., Saad I., Kassel G., Gargouri F., 2016. A core ontology of know-how and knowing-that for improving knowledge sharing and decision making in the digital age. *Journal of Decision Systems*, 26(2), 138–151. <https://doi.org/10.1080/12460125.2016.1252231>
- Godet M., 2008. Strategic foresight: Use and misuse of scenario building. *LIPSOR Working Paper*, 10, 143. <https://prospectivayfuturo.files.wordpress.com/2017/09/godet-strategicforesight-n10-2009.pdf>
- Godet M., Durance P., 2011. *La prospective stratégique – Pour les entreprises et les territoires* (2nd ed.), Dunod, 224 p.
- Gordon T.F., Prakken H., Walton D., 2007. The Carneades model of argument and burden of proof. *Artificial Intelligence*, 171(10–15), 875–896. <https://doi.org/10.1016/j.artint.2007.04.010>
- Gruber T.R., 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2), 199–220. <https://doi.org/10.1006/knac.1993.1008>

- Guillard V., Buche P., Destercke S., Tamani N., Croitoru M., Menut L. et al., 2015. A decision support system to design modified atmosphere packaging for fresh produce based on a bipolar flexible querying approach. *Computers and Electronics in Agriculture*, 111, 131–139. <https://doi.org/10.1016/j.compag.2014.12.010>
- Guillard V., Buche P., Dibie J., Dervaux S., Acerbi F., Chaix E. et al., 2016. CO₂ and O₂ solubility and diffusivity data in food products stored in data warehouse structured by ontology. *Data in Brief*, 7, 1556–1559. <https://doi.org/10.1016/j.dib.2016.04.044>
- Guillard V., Buche P., Menut L., Dervaux S., 2018. Matter Transfer Ontology. Recherche Data Gouv, V3. <https://doi.org/10.15454/NK24ID>
- Gwet K.L., 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics, LLC, 462 p.
- Hachem H., Chaib R.L., Engel E., Albert I., 2025. Ex-ante assessment of sample pooling implementation on the dietary exposure to nDL-PCB in meat. *Food Research International*, 214, 116599. <https://doi.org/10.1016/j.foodres.2025.116599>
- Handayati Y., Simatupang T.M., Perdana T., 2015. Agri-food supply chain coordination: the state-of-the-art and recent developments. *Logistics Research*, 8(1), 5. <https://doi.org/10.1007/s12159-015-0125-4>
- Hao T., We Y., Qiang J., Wang H., Lee K., 2017. The representation and extraction of quantitative information. *Proceedings of the 13th joint ISO-ACL workshop on interoperable semantic annotation (ISA-13)*. <https://aclanthology.org/W17-7408/>
- Hararay F., Norman R.Z., Cartwright D., 1965. Structural models: an introduction to the theory of directed graphs. Wiley, 439 p.
- Harispe S., 2014. Knowledge-based semantic measures: From theory to applications [Doctoral dissertation, Université de Montpellier]. <https://hal.science/tel-01175611v1>
- Health Canada, 2024. Nutrition labelling—Table of reference amounts for food—Canada. Government of Canada. <https://www.canada.ca/en/health-canada/services/technical-documents-labelling-requirements/nutrition-labelling-table-reference-amounts-food.html>
- Hecham A., Croitoru M., Bisquert P., 2017. Argumentation-based defeasible reasoning for existential rules. *AAMAS 2017 – 16th International Conference on Autonomous Agents and MultiAgent Systems*, 1568–1569. <https://hal.science/lirmm-01520171v1>
- Hecham A., Bisquert P., Croitoru M., 2018. On a flexible representation for defeasible reasoning variants. *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '18*, 1123–1131. <https://hal.science/lirmm-01894742v1>
- Hecham A., Bisquert P., Croitoru M., 2020a. A formalism unifying Defeasible Logics and Repair Semantics for existential rules. *ICCS 2020 - 25th International Conference on Conceptual Structures*, Online, Italy, 3–17. <https://hal.science/lirmm-03037374v1>
- Hecham A., Croitoru M., Bisquert P., 2020b. DAMN: Defeasible Reasoning Tool for Multi-Agent Reasoning. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, Association for the Advancement of Artificial Intelligence, New York, United States, 13612–13613. <https://hal.science/lirmm-02393877v1>
- Hignette G., Buche P., Dibie-Barthélemy J., Haemmerlé O., 2009. Fuzzy annotation of web data tables driven by a domain ontology. In: L. Aroyo, P. Traverso, F. Ciravegna et al. (Eds.), *The Semantic Web: Research and Application*: Vol. 5554. *Lecture Notes in Computer Science*, 638–653, Springer. https://doi.org/10.1007/978-3-642-02121-3_47
- Hitzler P., Krotzsch M., Rudolph S., 2009. Foundations of semantic web technologies. Chapman and Hall/CRC, 456 p. <https://doi.org/10.1201/9781420090512>
- Hofmann K., Tsagkias M., Meij E., De Rijke M., 2009. The impact of document structure on keyphrase extraction. *Proceedings of the 18th ACM conference on Information and knowledge management*, 1725–1728. <https://doi.org/10.1145/1645953.1646215>
- Hu G., Ahmed M., L'Abbé M.R., 2023. Natural language processing and machine learning approaches for food categorization and nutrition quality prediction compared with traditional methods. *The American Journal of Clinical Nutrition*, 117(3), 553–563. <https://doi.org/10.1016/j.ajcnut.2022.11.022>

- Huijbregts M.A.J., Steinmann Z.J.N., Elshout P.M.F., Stam G., Verones F., Vieira M. et al., 2017. ReCiPe2016: a harmonized life cycle impact assessment method at midpoint and endpoint level. *International Journal of Life Cycle Assessment*, 22, 138–147. <https://doi.org/10.1007/s11367-016-1246-y>
- Hwang C.-L., Yoon K., 1981. *Multiple attribute decision making. Methods and Applications – A state-of-the-art survey* (Vol. 186). Springer, 269 p. <https://doi.org/10.1007/978-3-642-48318-9>
- Ibanescu L., Dibie J., Dervaux S., Guichard E., Raad J., 2016. PO² – a process and observation ontology in food science. Application to dairy gels. In: E. Garoufallou, I. Subirats Coll, A. Stellato, J. Greenberg (Eds.), *Metadata and Semantics Research*, 672, 155–165, Springer. https://doi.org/10.1007/978-3-319-49157-8_13
- Ireland J.D., Möller A., 2010. LanguaL food description: a learning process. *European Journal of Clinical Nutrition*, 64, S44–S48. <https://doi.org/10.1038/ejcn.2010.209>
- Ireland J.D., Möller A., 2016. Food classification and description. *Encyclopedia of Food and Health*, 1–6. <https://doi.org/10.1016/B978-0-12-384947-2.00307-X>
- ISO (International Organization for Standardization), 2011. Information and documentation—Thesauri and interoperability with other vocabularies—Part 1: Thesauri for information retrieval (ISO Standard 25964-1:2011). <https://www.iso.org/standard/53657.html>
- ISO (International Organization for Standardization), 2013. Information and documentation—Thesauri and interoperability with other vocabularies—Part 2: Interoperability with other vocabularies (ISO Standard 25964-2:2011). <https://www.iso.org/standard/53658.html>
- Ispirova G., Eftimov T., Koroušić Seljak B., Korošec P., 2017. Mapping food composition data from various data sources to a domain-specific ontology. *Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, OIC3K*, 203–210. <https://doi.org/10.5220/0006504302030210>
- Jaccard P., 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 241–272.
- Jackson R., Matentzoglou N., Overton J.A., Vita R., Balhoff J.P., Buttigieg P.L. et al., 2021. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. *Database*, 2021, baab069. <https://doi.org/10.1093/database/baab069>
- Jain A., Cucerzan S., Azzam S., 2007. Acronym-expansion recognition and ranking on the web. *2007 IEEE International Conference on Information Reuse and Integration*, 209–214. <https://doi.org/10.1109/IRI.2007.4296622>
- Janowicz K., Haller A., Cox S.J.D., Le Phuoc D., Lefrançois M., 2019. SOSA: a lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics*, 56, 1–10. <https://doi.org/10.1016/j.websem.2018.06.003>
- Jessop D.M., Adams S.E., Murray-Rust P., 2011a. Mining chemical information from open patents. *Journal of Cheminformatics*, 3, 4. <https://doi.org/10.1186/1758-2946-3-4>
- Jessop D.M., Adams S.E., Willighagen E.L., Hawizy L., Murray-Rust P., 2011b. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 3, 41, 1–12. <https://doi.org/10.1186/1758-2946-3-41>
- Jones N., Bartlett H.E., Cooke R., 2019. An analysis of the impact of visual impairment on activities of daily living and vision-related quality of life in a visually impaired adult population. *British Journal of Visual Impairment*, 37(1), 50–63. <https://doi.org/10.1177/0264619618814071>
- Jørgensen C.E., Abrahamsen R.K., Rukke E.-O., Johansen A.-G., Schüller R.B., Skeie S.B., 2016. Optimization of protein fractionation by skim milk microfiltration: choice of ceramic membrane pore size and filtration temperature. *Journal of Dairy Science*, 99(8), 6164–6179. <https://doi.org/10.3168/jds.2016-11090>
- Kando N., 1997. Text-level structure of research papers: implications for text-based information processing systems. BCS-IRSG Annual Colloquium on IR Research.
- Kansou K., Chiron H., Della Valle G., Ndiaye A., Roussel P., 2014. Predicting the quality of wheat flour dough at mixing using an expert system. *Food Research International*, 64, 772–782. <https://doi.org/10.1016/j.foodres.2014.08.007>

- Kießling W., Köstler G., 2002. Preference SQL – Design, implementation, experiences. *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, 990–1001, Elsevier. <https://doi.org/10.1016/B978-155860869-6/50098-6>
- Kim D.H., Cho W.I., Lee S.J., 2020. Fault tree analysis as a quantitative hazard analysis with a novel method for estimating the fault probability of microbial contamination: a model food case study. *Food Control*, 110, 107019. <https://doi.org/10.1016/j.foodcont.2019.107019>
- Kommineni V.K., König-Ries B., Samuel S., 2024. From human experts to machines: an LLM supported approach to ontology and knowledge graph construction. *arXiv*. <https://doi.org/10.48550/arXiv.2403.08345>
- Konak A., Coit D.W., Smith A.E., 2006. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineering & System Safety*, 91(9), 992–1007. <https://doi.org/10.1016/j.res.2005.11.018>
- Kurtz A., Thomopoulos R., 2021. Safety vs. sustainability concerns of infant food users: French results and European perspectives. *Sustainability*, 13(18), 10074. <https://doi.org/10.3390/su131810074>
- La Fuente Arias C.I., González-Martínez C., Chiralt A., 2022. Lamination of starch/polyesters by thermocompression for food packaging purposes. *Research Square*. <https://doi.org/10.21203/rs.3.rs-2070061/v1>
- Lentschat M., Buche P., Dibie-Barthelemy J., Roche M., 2020. SciPuRe: a new representation of textual data for entity identification from scientific publications. *WIMS 2020: Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, 220–226. <https://doi.org/10.1145/3405962.34059>
- Lentschat M., 2021. TRANSMAT n-Ary relations [Dataset]. CIRAD Dataverse, V1. <https://doi.org/10.18167/DVN1/1BBJBQ>
- Lentschat M., Buche P., Dibie-Barthelemy J., Menut L., Roche M., 2021a. Food packaging permeability and composition dataset dedicated to text-mining. *Data in Brief*, 36, 107135. <https://doi.org/10.1016/j.dib.2021.107135>
- Lentschat M., Buche P., Menut L., 2021b. TRANSMAT Gold Standard. CIRAD Dataverse, V3. <https://doi.org/10.18167/DVN1/U7HK8J>
- Lentschat M., Buche P., Menut L., Guari R., 2021c. TRANSMAT tables data [Dataset]. CIRAD Dataverse, V3. <https://doi.org/10.18167/DVN1/GCZBC9>
- Lentschat M., Buche P., Dibie-Barthelemy J., Roche M., 2022a. A new method to extract n-Ary relation instances from scientific documents. *Expert Systems with Applications*, 209, 118332. <https://doi.org/10.1016/j.eswa.2022.118332>
- Lentschat M., Buche P., Dibie-Barthelemy J., Roche M., 2022b. Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications. *International Journal of Intelligent Information and Database Systems*, 15(1), 78–103. <https://doi.org/10.1504/IJIDS.2021.10042229>
- Lentschat M., Buche P., Menut L., Guari R., Roche M., 2022c. Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables. *Data in Brief*, 41, 108000. <https://doi.org/10.1016/j.dib.2022.108000>
- Liu J., Chabot Y., Troncy R., Huynh V.-P., Labbé T., Monnin P., 2023. From tabular data to knowledge graphs: a survey of semantic table interpretation tasks and methods. *Journal of Web Semantics*, 76, 100761. <https://doi.org/10.1016/j.websem.2022.100761>
- Lousteau-Cazalet C., Barakat A., Belaud J.-P., Buche P., Busset G., Charnomordic B. et al., 2016. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, 127, 351–367. <https://doi.org/10.1016/j.compag.2016.06.020>
- Lundberg S., Lee S.-I., 2017. A unified approach to interpreting model predictions. *arXiv*. <https://doi.org/10.48550/arXiv.1705.07874>
- Maedche A., Staab S., 2000. Mining ontologies from text. In: R. Dieng, O. Corby (Eds.), *Knowledge engineering and knowledge management methods, models, and tools*, 189–202, Springer. https://doi.org/10.1007/3-540-39967-4_14

- Mahajan P.V., Oliveira F.A.R., Montanez J.C., Frias J., 2007. Development of user-friendly software for design of modified atmosphere packaging for fresh and fresh-cut produce. *Innovative Food Science & Emerging Technologies*, 8(1), 84–92. <https://doi.org/10.1016/j.ifset.2006.07.005>
- Manning C.D., Raghavan P., Schütze H., 2008. Evaluation in information retrieval. *In: Introduction to information retrieval*, 188–210, Cambridge University Press. <https://doi.org/10.1017/CBO9780511809071>
- Mao S., Rosenfeld A., Kanungo T., 2003. Document structure analysis algorithms: a literature survey. *In: T. Kanugo, E.H. Barney Smith, J. Hu, P.B. Kantor (Eds.), Proceedings SPIE 5010, Document Recognition and Retrieval X*, 197–207. <https://doi.org/10.1117/12.476326>
- Matentzoglou N., Balhoff J.P., Bello S.M., Bizon C., Brush M., Callahan T.J. et al., 2022. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database, 2022, baac035. <https://doi.org/10.1093/database/baac035>
- Mavani N.R., Ali J.M., Othman S., Hussain M.A., Hashim H., Rahman N.A., 2022. Application of artificial intelligence in food industry—a guideline. *Food Engineering Reviews*, 14, 134–175. <https://doi.org/10.1007/s12393-021-09290-z>
- Mayer-Laigle C., Beaugrand J., Bourmaud A., Brionne L., Colinart T., Dervaux S. et al., 2024. Datasets on the production routes and the properties of plant powders for manufacturing of green products. *Data in Brief*, 56, 110787. <https://doi.org/10.1016/j.dib.2024.110787>
- Menard C., Dumas C., Goglia R., Spiteri M., Gillot N., Combris P. et al., 2011. OQALI: a French database on processed foods. *Journal of Food Composition and Analysis*, 24(4–5), 744–749. <https://doi.org/10.1016/j.jfca.2010.09.001>
- Mermet L., 2004. Prospective : Un objet d'étude pour les SIC. *Hermès, La Revue*, 38(1), 207–214. <https://doi.org/10.4267/2042/9451>
- Mitchell R.K., Agle B.R., Wood D.J., 1997. Toward a theory of stakeholder identification and salience: defining the principle of who and what really counts. *The Academy of Management Review*, 22(4), 853–886. <https://doi.org/10.2307/259247>
- Modgil S., Prakken H., 2014. The ASPIC⁺ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1). <https://doi.org/10.1080/19462166.2013.869766>
- Muljarto A.-R., Salmon J.-M., Neveu P., Charnomordic B., Buche P., 2014. Ontology-based model for food transformation processes – Application to winemaking. *In: S. Closs, R. Studer, E. Garoufallou, M.-A. Sicilia (Eds.), Metadata and Semantics Research – MTSR 2014. Communications in Computer and Information Science*, 478, 329–343. https://doi.org/10.1007/978-3-319-13674-5_30
- Munch M., Buche P., Manfredotti C., Wuillemin P.-H., Angellier-Coussy H., 2021. A process reverse engineering approach using Process and Observation Ontology and probabilistic relational models: application to processing of bio-composites for food packaging. *In: E. Garoufallou, M.-A. Ovalle-Perandones, A. Vlachidis (Eds.), Metadata and Semantics Research – MTSR 2021 – 15th International Conference on Metadata and Semantics Research*, 1537, 3–15, Springer. https://doi.org/10.1007/978-3-030-98876-0_1
- Munch M., Buche P., Angellier-Coussy H., Manfredotti C., Wuillemin P.-H., 2022a. Formalising contextual expert knowledge for causal discovery in linked knowledge graphs about transformation processes: application to processing of bio-composites for food packaging. *International Journal of Metadata, Semantics and Ontologies*, 16(1), 1–15. <https://doi.org/10.1504/IJMSO.2022.131129>
- Munch M., Buche P., Dervaux S., Breyse A., Berthet M.-A., David G. et al., 2022b. Biocomposites from poly(3-hydroxybutyrate-co-3-hydroxyvalerate) and lignocellulosic fillers: processes stored in data warehouse structured by an ontology. *Data in Brief*, 42, 108191. <https://doi.org/10.1016/j.dib.2022.108191>
- Munch M., Buche P., Dervaux S., Dibie J., Ibanescu L., Manfredotti C. et al., 2022c. Combining ontology and probabilistic models for the design of bio-based product transformation processes. *Expert Systems with Applications*, 203, 117406. <https://doi.org/10.1016/j.eswa.2022.117406>
- Munch M., Buche P., Menut L., Cufi J., Guillard V., 2023a. CO₂ solubility and composition data of food products stored in data warehouse structured by an ontology. *Data in Brief*, 47, 108950. <https://doi.org/10.1016/j.dib.2023.108950>
- Munch M., Guillard V., Gaucel S., Destercke S., Thévenot J., Buche P., 2023b. Composition-based statistical model for predicting CO₂ solubility in modified atmosphere packaging application. *Journal of Food Engineering*, 340, 111283. <https://doi.org/10.1016/j.jfoodeng.2022.111283>

- Munch M., Rezette L., Buche P., Chambrey B., Deborde C., Dervaux S. et al., 2025. Dataset for common wheat (*Triticum aestivum* L.) grain and flour characterization using classical and advanced analyses. *Data in Brief*, 59, 111375. <https://doi.org/10.1016/j.dib.2025.111375>
- Navarro-Dols J., González-Pernía J.L., 2020. Gastronomy as a real agent of social change. *International Journal of Gastronomy and Food Science*, 21, 100240. <https://doi.org/10.1016/j.ijgfs.2020.100240>
- Navarro-Dols J., González-Pernía J.L., Georgantzis N., Ashta A., 2022. Team profiles and roles in Social Gastronomy: a qualitative study in Spain. *International Journal of Gastronomy and Food Science*, 27, 100457. <https://doi.org/10.1016/j.ijgfs.2021.100457>
- Nebot V., Berlanga R., 2009. Building tailored ontologies from very large knowledge resources. *Proceedings of the 11th International Conference on Enterprise Information – Volume 2: ICEIS*, 144–151. <https://doi.org/10.5220/0001984001440151>
- Nguyen P., Kertkeidkachorn N., Ichise R., Takeda H., 2019. MTab: matching tabular data to knowledge graph using probability models. *arXiv*. <http://arxiv.org/abs/1910.00246>
- OECD (Organisation for Economic Co-operation and Development), 2021. Food systems and the triple challenge, No. 2021/01, OECD Publishing, Paris. <https://doi.org/10.1787/2b4fea77-en>
- Okazaki N., Ananiadou S., 2006. A term recognition approach to acronym recognition. COLING-ACL '06: Proceedings of the COLING/ACL on Main conference poster sessions, 643–650.
- Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C.L., Mishkin P. et al., 2022. Training language models to follow instructions with human feedback. *arXiv*. <https://doi.org/10.48550/arXiv.2203.02155>
- Paas F., Tuovinen J.E., Tabbers H., Van Gerven P.W.M., 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71. https://doi.org/10.1207/S15326985EP3801_8
- Pagels H.R., Dreyfus H.L., McCarthy, J., Minsky M.L., Papert S., Searle J., 1984. Panel discussion: has artificial intelligence research illuminated human thinking? *Annals of the New York Academy of Sciences*, 426(1), 138–160. <https://doi.org/10.1111/j.1749-6632.1984.tb16517.x>
- Pahasup-anan T., Kreetachat T., Ruengphrathuengsuka, W., Wongcharee S., Usahanunth N., Imman S., Suwannahong K., 2022. Dust explosion risk assessment of extruded food production process by fault tree analysis. *ACS Chemical Health & Safety*, 29(1), 91–97. <https://doi.org/10.1021/acs.chas.1c00036>
- Pakkala H., Christensen T., Martínez De Victoria I., Presser K., Kadvan A., 2010. Harmonised information exchange between decentralised food composition database systems. *European Journal of Clinical Nutrition*, 64(S3), S58–S63. <https://doi.org/10.1038/ejcn.2010.212>
- Pearl J., 2009. Causality. Models, reasoning, and inference (2nd ed.). Cambridge University Press, 465 p. <https://doi.org/10.1017/CBO9780511803161>
- Pénicaud C., Ibanescu L., Allard T., Fonseca F., Dervaux S., Perret B. et al., 2019. Relating transformation process, eco-design, composition and sensory quality in cheeses using PO² ontology. *International Dairy Journal*, 92, 1–10. <https://doi.org/10.1016/j.idairyj.2019.01.003>
- Perrignon M., Emily M., Munch M., Jeantet R., Croguennec T., 2025. Machine learning for predicting industrial performance: example of the dry matter content of emmental-type cheese. *International Dairy Journal*, 162, 106143. <https://doi.org/10.1016/j.idairyj.2024.106143>
- Perrignon M., Emily M., Munch M., Debuire P., Jeantet R., Croguennec T., 2026. A digital twin integrating multi-objective optimization to support fryer operators in managing potato crisps production. *Journal of Food Engineering*, 406, 112800. <https://doi.org/10.1016/j.jfoodeng.2025.112800>
- Pesonen H.-L., Ekvall T., Fleischer G., Hupples G., Jahn C., Klos Z.S. et al., 2000. Framework for scenario development in LCA. *The International Journal of Life Cycle Assessment*, 5(1), 21–30. <https://doi.org/10.1007/bf02978555>
- Pour M.A.N., Algergawy A., Buche P., Castro L.J., Chen J., Dong H. et al., 2022. Results of the Ontology Alignment Evaluation Initiative 2022. *CEUR Workshop Proceedings. 17th International Workshop on Ontology Matching (OM 2022)*. <https://openaccess.city.ac.uk/id/eprint/29879/>
- Pour M.A.N., Algergawy A., Buche P., Castro L.J., Chen J., Coulet A. et al., 2023. Results of the Ontology Alignment Evaluation Initiative 2023. *Proceedings of the 18th International Workshop on Ontology Matching Co-Located with the 22nd International Semantic Web Conference (ISWC 2023)*, 3591, 97–139. <https://ceur-ws.org/Vol-3591/>

- Pour M.A.N., Algergawy A., Blomqvist E., Buche P., Chen J., Cotovio P.G. et al., 2024. Results of the Ontology Alignment Evaluation Initiative 2024. In: E. Jiménez-Ruiz, O. Hassanzadeh, C. Trojahn et al. (Eds.), *Proceedings of the 19th International Workshop on Ontology Matching (OM 2024)*, 64–97. <https://doi.org/10.5445/IR/1000179469>
- Poveda-Villalón M., Fernández-Izquierdo A., Fernández-López M., García-Castro R., 2022. LOT: an industrial oriented ontology engineering framework. *Engineering Applications of Artificial Intelligence*, 111, 104755. <https://doi.org/10.1016/j.engappai.2022.104755>
- Pustejovsky J., Castano J., Cochran B., Kotecki M., Morrell M., 2001. Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Studies in Health Technology and Informatics*, 371–375.
- Qi P., Zhang Y., Zhang Y., Bolton J., Manning C.D., 2020. Stanza: a Python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. <https://doi.org/10.18653/v1/2020.acl-demos.14>
- Rayaleh E., Ang M., 2024. Saveurs évoluées, Agricultural engineering finalized project. Institut Agro-Montpellier.
- Rijgersberg H., Wigham M., Top J.L., 2011. How semantics can improve engineering processes: a case of units of measure and quantities. *Advanced Engineering Informatics*, 25(2), 276–287. <https://doi.org/10.1016/j.aei.2010.07.008>
- Rijgersberg H., van Assem M., Top J., 2013. Ontology of units of measure and related concepts. *Semantic Web*, 4(1), 3–13.
- Ru C., Tang J., Li S., Xie S., Wang T., 2018. Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing & Management*, 54(4), 593–608. <https://doi.org/10.1016/j.ipm.2018.04.002>
- Saint-Dizier P., 2012. Processing natural language arguments with the <TextCoop> platform. *Argument & Computation*, 3(1), 49–82. <https://doi.org/10.1080/19462166.2012.663539>
- Salton G., McGill, M.J., 1986. *Introduction to modern information retrieval*. McGraw-Hill, Inc, 400 p.
- Sandjong-Sayon D., Cancalon M., Kurtz A., Meurillon M., Bourliou Lacanal C., Engel E., Thomopoulos R., 2022. Aliments infantiles : Perceptions des utilisateurs et nouvelles questions de recherche. Journées Francophones de Nutrition (JFN 2022). <https://hal.inrae.fr/hal-03856714>
- Shah P.K., Perez-Iratxeta C., Bork P., Andrade M.A., 2003. Information extraction from full text scientific articles: where are the keywords? *BMC bioinformatics*, 4, 20. <https://doi.org/10.1186/1471-2105-4-20>
- Shvaiko P., Euzenat J., 2013. Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1), 158–176. <https://doi.org/10.1109/TKDE.2011.253>
- Soldatova L.N., Nadis D., King R.D., Basu P.S., Haddi E., Baumlé V. et al., 2014. EXACT2: the semantics of biomedical protocols. *BMC Bioinformatics*, 15(S14), S5. <https://doi.org/10.1186/1471-2105-15-S14-S5>
- Sohn J., Bisquert P., Buche P., Hecham A., Kalbar P.P., Goldstein B. et al., 2020. Argumentation corrected context weighting-life cycle assessment: a practical method of including stakeholder perspectives in multi-criteria decision support for LCA. *Sustainability*, 12(6), 2170. <https://doi.org/10.3390/su12062170>
- Sousa-Gallagher M.J., Mahajan P.V., 2013. Integrative mathematical modelling for MAP design of fresh-produce: theoretical analysis and experimental validation. *Food Control*, 29(2), 444–450. <https://doi.org/10.1016/j.foodcont.2012.05.072>
- Spirtes P., Glymour C.N., Scheines R., 2001. *Causation, prediction, and search* (2nd ed). MIT Press. <https://doi.org/10.7551/mitpress/1754.001.0001>
- Studer R., Benjamins V.R., Fensel D., 1998. Knowledge engineering: principles and methods. *Data & Knowledge Engineering*, 25(1–2), 161–197. [https://doi.org/10.1016/S0169-023X\(97\)00056-6](https://doi.org/10.1016/S0169-023X(97)00056-6)
- Subhasri D., Dutta S. Maria Leena M., Moses J.A., Anandharamkrishnan, C., 2022. Gastronomy: an extended platform for customized nutrition. *Future Foods*, 5, 100147. <https://doi.org/10.1016/j.fufo.2022.100147>
- Sun A., 2009. A two-stage bootstrapping algorithm for relation extraction. *Proceedings of the Student Research Workshop*, 76–82.
- Taillandier F., Guignard S., Pares N., Curt C., Schleyer-Lindenmann A., 2023. Analysing debate on infrastructures using abstract argumentation. *Urban Governance*, 3(2), 138–147. <https://doi.org/10.1016/j.ugj.2023.02.004>

- Thomopoulos R., Croitoru M., Tamani N., 2015. Decision support for agri-food chains: a reverse engineering argumentation-based approach. *Ecological Informatics*, 26(Part 2), 182–191. <https://doi.org/10.1016/j.ecoinf.2014.05.010>
- Thomopoulos R., 2018. A practical application approach to argumentation for multicriteria analysis and decision support. *EURO Journal on Decision Processes*, 6(3–4), 237–255. <https://doi.org/10.1007/s40070-018-0087-2>
- Thomopoulos R., Baudrit C., Boukhelifa N., Boutrou R., Buche P., Guichard E. et al., 2019. Multi-criteria reverse engineering for food: genesis and ongoing advances. *Food Engineering Reviews*, 11, 44–60. <https://doi.org/10.1007/s12393-018-9186-x>
- Thomopoulos R., Cufi J., Le Breton M., 2020. A generic software to support collective decision in food chains and in multi-stakeholder situations. FoodSim 2020 – 11th Biennial FOODSIM Conference, Ghent, Belgium. <https://hal.science/hal-02484363>
- Thomopoulos R., Chaib R.L., Dervilly G., 2025. Benefit-risk analysis of prospective scenarios implementing new-generation food safety management technologies: Qualitative, stakeholder-driven study of the infant food chain. *Food Research International*, 214, 116687. <https://doi.org/10.1016/j.foodres.2025.116687>
- Tigrine A.N., Bellahsene Z., Todorov K., 2016. Selecting optimal background knowledge sources for the Ontology Matching Task. In: E. Blomqvist, P. Ciancarini, F. Poggi, F. Vitali (Eds.), Knowledge Engineering and Knowledge Management – EKAW 2016. *Lecture Notes in Computer Science*, 10024, 651–665, Springer. https://doi.org/10.1007/978-3-319-49004-5_42
- Tjong Kim Sang E.F., De Meulder F., 2003. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *arXiv*. <https://doi.org/10.48550/arXiv.cs/0306050>
- Toro S., Anagnostopoulos A., Bello S., Blumberg K., Cameron R., Carmody L. et al., 2024. Dynamic rRetrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *Journal of Biomedical Semantics*, 15(1), 19. <https://doi.org/10.1186/s13326-024-00320-3>
- Touhami R., Buche P., Dibie-Barthélemy J., Ibănescu L., 2011. An ontological and terminological resource for n-ary relation annotation in web data tables. In: R. Meersman, T. Dillon, P. Herrero et al. (Eds.), On the Move to Meaningful Internet Systems: OTM 2011. *Lecture Notes in Computer Science*, 7045, 662–679, Springer. https://doi.org/10.1007/978-3-642-25106-1_19
- Truswell A.S., Bateson D.J., Madafiglio K.C., Pennington J.A.T., Rand W.M., Klensin J.C., 1991. INFOODS Guidelines for Describing Foods: a systematic approach to describing foods to facilitate international exchange of food composition data. *Journal of Food Composition and Analysis*, 4(1), 18–38. [https://doi.org/10.1016/0889-1575\(91\)90045-8](https://doi.org/10.1016/0889-1575(91)90045-8)
- Uschold M., Gruninger M., 1996. Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11(2), 93–136. <https://doi.org/10.1017/S0269888900007797>
- Van Assem M., Rijgersberg H., Wigham M., Top J., 2010. Converting and annotating quantitative data tables. In: P.F. Patel-Schneider, Y. Pan, P. Hitzler et al. (Eds). The Semantic Web – ISWC 2010. *Lecture Notes in Computer Science*, 6496, 16–31, Springer. https://doi.org/10.1007/978-3-642-17746-0_2
- Van Hage W.R., Sini M., Finch L., Kolb H., Schreiber G., 2010. The OAEI food task: an analysis of a thesaurus alignment task. *Applied Ontology*, 5(1), 1–28. <https://doi.org/10.3233/AO-2010-0072>
- Vilone G., Longo L., 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89–106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- Vivas C., 2022. Étude des bénéfices et risques de l'entomoconversion de biodéchets urbains par *Hermetia illucens*. (Version 907) [Master 1 report]. Montpellier university.
- Vivas C., Planche C., Macombe C., Borel P., Engel E., Thomopoulos R., 2022. Une analyse systémique et interdisciplinaire de scénarios de développement durable basés sur l'entomoconversion. *Revue Francophone du développement durable*, 20. <https://hal.inrae.fr/hal-03999567>
- Vuillet C., Guillard V., Angellier-Coussy H., Sudre G., Gouanvé F., Fleury E., Charlot A., 2026. Hydrophobization of carboxymethyl cellulose by Passerini reaction: towards films with improved water vapor barrier properties. *Journal of Membrane Science*, 738, 124851. <https://dx.doi.org/10.1016/j.memsci.2025.124851>

- Wang D., Zhang H., 2010. Inverse-category-frequency based supervised term weighting scheme for text categorization. *arXiv*. <https://doi.org/10.48550/arXiv.1012.2609>
- Wang S., Sun X., Li X., Ouyang R., Wu F., Zhang T. et al., 2023. GPT-NER: Named entity recognition via large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2304.10428>
- Weber M., Buche P., Ibanescu L., Dervaux S., Guillemin H., Cufi J. et al., 2023. PO2/TransformON, an ontology for data integration on food, feed, bioproducts and biowaste engineering. *npj Science of Food*, 7(1), 47. <https://doi.org/10.1038/s41538-023-00221-2>
- Wolf C., Angellier-Coussy H., Gontard N., Doghieri F., Guillard V., 2018. How the shape of fillers affects the barrier properties of polymer/non-porous particles nanocomposites: a review. *Journal of Membrane Science*, 556, 393–418. <https://doi.org/10.1016/j.memsci.2018.03.085>
- Wood I., Johnson M., Wan S., 2021. Integrating lexical information into entity neighbourhood representations for relation prediction. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3429–3436. <https://doi.org/10.18653/v1/2021.naacl-main.268>
- Wren J.D., Garner H.R., 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of information in medicine*, 41(05), 426–434. <https://doi.org/10.1055/s-0038-1634373>
- Xu J., Huang Y.-L., 2005. A machine learning approach to recognizing acronyms and their expansion. *2005 International Conference on Machine Learning and Cybernetics*, 4, 2313–2319. <https://doi.org/10.1109/ICMLC.2005.1527330>
- Yang G., Kifer M., Zhao C., 2003. Flora-2: a rule-based knowledge representation and inference infrastructure for the Semantic Web. In: R. Meersman, Z. Tari, D.C. Schmidt (Eds.), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, 2888, Springer. https://doi.org/10.1007/978-3-540-39964-3_43
- Yates A., Banko M., Broadhead M., Cafarella M.J., Etzioni O., Soderland S., 2007. Textrunner: Open information extraction on the web. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 25–26. <https://aclanthology.org/N07-4013/>
- Yun B., Buche P., Bisquert P., Costa S., Croitoru M., Cufi J. et al., 2018. Consumer perception data and scientific arguments about food packaging functionalities for fresh strawberries. *Data in Brief*, 20, 1924–1927. <https://doi.org/10.1016/j.dib.2018.09.034>
- Zadeh L.A., 1965. Fuzzy sets. *Information and Control*, 8(3), 338–353. [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
- Zaenen A., Condoravdi C., Bobrow D., Hoffmann R., 2010. Supporting rule-based representations with corpus-derived lexical information. *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, 114–121. <https://aclanthology.org/W10-0914/>
- Zeng M.L., 2017. Smart data for digital humanities. *Journal of Data and Information Science*, 2(1), 1–12. <https://doi.org/10.1515/jdis-2017-0001>
- Zhang Z., 2017. Effective and efficient semantic table interpretation using TableMiner⁺. *Semantic Web – Interoperability, Usability, Applicability*, 8(6), 921–957. <https://doi.org/10.3233/SW-160242>
- Zhou D., Zhong D., He Y., 2014. Biomedical relation extraction: From binary to complex. *Computational and Mathematical Methods in Medicine*, 2014. <https://doi.org/10.1155/2014/298473>
- Zurayk R., Ghya T., 2009. The Lebanese terroir: a challenge of quality. In: *Localizing Products: A sustainable approach for natural and cultural diversity in the south?* International symposium.

Acknowledgments

The research presented here by computer scientists specialized in knowledge engineering is the result of long-term collaboration with many agricultural resource processing experts from the INRAE TRANSFORM Division and associated partners. The authors of this book would like to take this opportunity to thank them for their support in helping us design the solutions proposed here.

We would like to extend special thanks to:

The INRAE TRANSFORM Division, for its support to the ICO Knowledge Engineering team and role as the lead division since 2019 in the INRAE CATI DIISCICO professional network of IT engineers. Many of this book's co-authors are members of both the division and the network.

The INRAE MATHNUM Division, for its unwavering support of the IN-OVIVE INRAE scientific exchange network for knowledge engineering, co-founded in 2012 by one of this book's editors. Many of this book's co-authors are also members of this network.

The INRIA BOREAL team, a long-term partner working closely with the ICO Knowledge Engineering team (several of this book's co-authors work on both teams), for the stimulating scientific environment it provides on knowledge representation and reasoning on heterogeneous data. We would especially like to thank Madalina Croitoru (now on the Interactive Digital Humans team of the Robotics Department at the Montpellier Laboratory of Computer Science, Robotics, and Microelectronics – LIRMM), and Abdelraouf Hecham, whose PhD was jointly supervised by Madalina Croitoru and one of this book's editors and gave rise to one of the approaches presented in this book.

Juliette Dibie (from INRAE's Human Resources Department), who, much to our regret, changed career paths in 2020, for her significant contribution to the design of several knowledge engineering solutions presented in several chapters. Juliette co-supervised several PhD theses dedicated to the design of two task ontologies (@Web and PO2) presented in Chapter 2, the coupling of knowledge engineering and natural language processing approaches presented in Chapter 3, and the coupling of knowledge engineering and machine learning models presented in Chapter 4. Juliette also participated in creating and managing the IN-OVIVE INRAE scientific exchange network with one of this book's editors.

Sébastien Destercke (from the HEUDIASYC Joint Research Unit), a specialist in imprecise probabilistic approaches, for his long-running collaboration that started when he was a member of the ICO team and has continued with his help in formalizing some aspects of the knowledge engineering approaches presented in this book.

Mathieu Roche (from the TETIS Joint Research Unit), a specialist in text mining, for his long-term support in the coupling of knowledge engineering and natural language processing approaches, such as the one presented in Chapter 3.

Brigitte Charnomordic (from the MISTEA Joint Research Unit), retired, a specialist of reasoning under uncertainty, for her long-term collaboration in designing decision support systems applied to biomass processing, such as the one presented in Chapter 4.

Pascal Neveu (from the MISTEA Joint Research Unit), recently retired, for the common vision on the importance of Semantic Web technologies in creating links between all INRAE fields of research and the common initiatives (IN-OVIVE INRAE scientific exchange network, INRAE Semantic Linked Data seminars, among others).

Valérie Guillard, Hélène Coussy, Stéphane Peyron, and Nathalie Gontard (from the EPOP team at the IATE Joint Research Unit), specialists in food packaging design; Claire Mayer (from the PHYPRODIV team at IATE); and Leslie Lhomond and Adrien Reau (from the PLANET platform at IATE), specialists in plant-based powder production and applications, for their long-standing collaboration. These colleagues enabled us to progressively improve the IT solutions proposed in this book, ensuring the complex application requirements were taken into account.

Jean-Pierre Belaud, Caroline Sablayrolles, Claire Vialle, and Rachid Ouaret (from LGC and LCA Joint Research Units), specialists in agricultural resource processing sustainability analyses that include environmental and economic aspects, for their long-standing collaboration. These colleagues helped us to continuously improve the IT solutions presented in this book in order to validate their relevance for sustainability analyses.

Caroline Pénicaud (from the SAYFOOD Joint Research Unit), a specialist in the eco-design of foods, bioproducts, and (bio-)processes. Caroline collaborated with authors of this book to develop the Competency Questions used to design the PO² task ontology. She also coordinated the ANR DataSusFood project, a key initiative in building the PO² ecosystem presented in this book and which received the Open Science Award from the French Ministry of Research, and she continues to influence its evolution.

Elisabeth Guichard (from the PAM Joint Research Unit), now retired, who also contributed to the design of the PO² task ontology.

Geneviève Gesan-Guiziou (from the STLO Joint Research Unit), a specialist in milk microfiltration who enabled us to challenge and improve the IT solutions proposed in this book on this complex application. Geneviève also coordinates the INRAE group of experts in multicriteria assessment methods to which one of this book's editors belongs. She gave us the opportunity to position our work at the scale of agrifood research.

Bernard Cuq (from the GRAINE.S team at the IATE Joint Research Unit), a specialist in cereal processing who gave us the opportunity to test our methods and tools in projects involving industrial partners with very positive results, and Abdellatif Barakat (from the IATE Joint Research Unit), a specialist in green chemistry who worked with us to design decision support tools to compare agricultural waste valorization pathways.

The ICO Knowledge Engineering team at IATE, and in particular its nonpermanent members (PhD students, postdoctoral researchers, engineers, and interns) who, at different times and to varying degrees, contributed to the design and implementation of the approaches presented in this book.

Finally, our warmest thanks go to everyone at Éditions Quæ for supporting us step by step throughout the publishing process. We are also very grateful to Teri Jones-Villeneuve, a translation and editing specialist, for her meticulous work in homogenizing the content of the book and adapting it for an international audience.



List of authors

Reine Barbar, Institut Agro Montpellier, UMR IATE, INRAE, University of Montpellier, Montpellier, France

Cédric Baudrit, INRAE, UMR I2M, Bordeaux University, Talence, France

Pierre Bisquert, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Patrice Buche, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Romy-Lynn Chaib, R.-L. Chaib Society, Pineuilh, France

Julien Cufi, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Julien Couteaux, Docamex, Talence, France

Stéphane Dervaux, INRAE, UMR MIA Paris-Saclay, Université Paris-Saclay, AgroParisTech, Palaiseau, France

Juliette Defossez, INRAE, UMR 1208 IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Erwan Engel, INRAE, UR QuaPA, Saint-Genès-Champanelle, France

Christophe Fernandez, INRAE, UMR I2M, Bordeaux University, Talence, France

Hervé Guillemain, INRAE, UMR PAM, Institut Agro Dijon, Université Bourgogne Europe, Poligny, France

Nicolas Guinet, INRAE, UMR PSAE, Université Paris-Saclay, AgroParisTech, Palaiseau, France

Yannick Guyonvarch, INRAE, UMR PSAE, Université Paris-Saclay, AgroParisTech, Palaiseau, France

Liliana Ibanescu, AgroParisTech, UMR MIA Paris-Saclay, Université Paris-Saclay, INRAE, 91120, Palaiseau, France

Martin Lentschat, Université de Toulouse 2 – Jean Jaurès, UMR LERASS, Université de Toulouse, Université de Montpellier 3 – Paul Valéry, Université de technologie de Tarbes Occitanie Pyrénées, Toulouse, France

Catherine Macombe, INRAE, UMR ITAP, Institut Agro Montpellier 34060 Montpellier, France

Luc Menut, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Mélanie Münch, INRAE, UMR 1253 STLO, Institut Agro Rennes, Rennes, France

Alrick Oudot, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Emmanuele Piaud, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Mathieu Roche, CIRAD, UMR TETIS, AgroParisTech, INRAE, Montpellier, France

Rallou Thomopoulos, INRAE, UMR IATE, University of Montpellier, Institut Agro Montpellier, Montpellier, France

Magalie Weber, INRAE, UR BIA, 44316 Nantes, France

Layout: Hélène Bonnet

Printed in France by

Print number:

Legal deposit:

This book presents workflows – combinations of digital processing steps based on knowledge engineering – to address complex problems in the agrifood and bioproduct industries. These workflows enable the collection, organization, sharing, linking, and updating of various sources: expert opinions, experimental data, and stakeholder preferences. The main methodological challenge is the extraction, representation, and integration of imperfect data for reusing in tasks such as learning, prediction, and decision support.

Knowledge engineering, a branch of AI, provides the methods and tools needed for expertise formalization, automated reasoning, and supported decision-making. The book builds on the work of the INRAE ICO team at the IATE research unit (Montpellier, France) and the INRAE CATI DIISCICO network, which were awarded the Open Science Research Data Award 2023 by the French Ministry of Research. This research results from a long-term collaboration with experts from the INRAE TRANSFORM division and their partners.

The workflows are implemented in innovative tools for capitalizing and processing data/knowledge; applied to specific agrifood sector issues; and intended to support collective decision-making. They are illustrated with concrete cases, such as the design of biodegradable packaging composites, the selection of food product packaging, and technological recommendations for cheese quality.

The book details the methodologies and tools based on the generic nature of the approaches to ensure their durability, independent of technology, while also providing open resources (software, datasets, ontologies, thesauri) so they can be reused or adapted to other contexts.

Patrice Buche is a Research Engineer at INRAE. He has a PhD in Computer Science and a qualified Senior Researcher diploma (HDR Paris-Saclay), and is an expert in the integration of heterogeneous data and multicriteria decision-support systems for food and bioproducts.

Pierre Bisquert is a Researcher at INRAE with a PhD in Computer Science (Paul Sabatier Univ., Toulouse). He is a specialist in knowledge representation and reasoning on heterogeneous and inconsistent data, argumentation theory, and collective decision-making.

Rallou Thomopoulos is a Senior Researcher at INRAE. She holds a PhD in Knowledge Engineering (AgroParisTech) and works on collective and explainable decision-making, multi-actor argumentation, opinion analysis, and ethics applied to agrifood chains.

€39

ISBN: 978-2-7592-4316-7



éditions
Quæ

Éditions CIRAD, Ifremer, INRAE
www.quae.com

INRAE

ISSN: 1777-4624
Ref.: 03076