frontiers | Research Topics

Knowledge graph technologies: The next Frontier of the food, agriculture, and water domains

Edited by
Marie-Angélique Laporte, Catherine Roussey and Christophe Guéret

Published in
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-4182-1
DOI 10.3389/978-2-8325-4182-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Knowledge graph technologies: The next Frontier of the food, agriculture, and water domains

Topic editors

Marie-Angélique Laporte — Bioversity International, France
Catherine Roussey — INRAE Occitanie Montpellier, France
Christophe Guéret — Accenture, Ireland

Citation

Laporte, M.-A., Roussey, C., Guéret, C., eds. (2023). *Knowledge graph technologies: The next Frontier of the food, agriculture, and water domains*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4182-1

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 28 November 2023
DOI 10.3389/frai.2023.1319844

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Matthew McCabe,
King Abdullah University of Science and Technology, Saudi Arabia

*CORRESPONDENCE
Catherine Roussey
✉ catherine.roussey@inrae.fr
Christophe Guéret
✉ christophe.gueret@accenture.com
Marie-Angélique Laporte
✉ m.a.laporte@cgiar.org

RECEIVED 11 October 2023
ACCEPTED 24 October 2023
PUBLISHED 28 November 2023

CITATION
Roussey C, Guéret C and Laporte M-A (2023) Editorial: Knowledge graph technologies: the next Frontier of the food, agriculture, and water domains. *Front. Artif. Intell.* 6:1319844.
doi: 10.3389/frai.2023.1319844

Editorial: Knowledge graph technologies: the next Frontier of the food, agriculture, and water domains

Catherine Roussey[1]*, Christophe Guéret[2]* and Marie-Angélique Laporte[3]*

[1]MISTEA, INRAE & Institut Agro, University of Montpellier, Montpellier, France, [2]Accenture, Technology Innovation Labs, The Dock, Dublin, Ireland, [3]Digital Inclusion, Bioversity International, Montpellier, France

KEYWORDS

agriculture, food, knowledge graph, ontology, graph database, Semantic Web, property graph, inference

Editorial on the Research Topic
Knowledge graph technologies: the next Frontier of the food, agriculture, and water domains

A Knowledge Graph (KG) is based on a graph model to encode the description of entities. As defined by Hogan and his collaborators in 2022, a knowledge graph is "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities." For Knowledge Graph using Semantic Web technologies, entities (people, events, concepts, etc.) are identified by a Uniform Resource Identifier (URI). This URI is the source of a graph description, the edge specifies the nature of the link (person name or brotherhood relationship) and the destination of the edge could be a simple literal (the person name) or a URI that identifies another entity (the URI of the brother). The main advantage of these technologies is to link entities that are described differently in several knowledge graphs provided by various organizations. Thus, computer scientists may analyze all those graph descriptions to derive new information (detect incoherencies, complete data, etc.).

During the last decade, considerable progress has been made in the construction and enrichment of KGs, including ontology matching, data integration, fact prediction, and validation. This happened largely thanks to the use of techniques developed in the fields of knowledge representation, reasoning, and machine learning. With these advances, more and more applications are now able to produce and process KGs in domains such as life sciences, Galleries/Libraries/Archives/Museums (GLAMs), and health care. The subjects of interest within the Food, Agriculture, and Water domains are often complex phenomena where entities evolve through time and space. Those phenomena may be transformed by different processes and influenced by both human and natural systems. The scientific disciplines that study these phenomena are diverse and do not necessarily share the same vocabularies, the same techniques of observation, the same analyses, and so on. Indeed, each discipline often has its own point of view to describe the complexity of the studied phenomena. KG technologies provide one possible approach to express this diversity of representations and align or combine them.

This Research Topic has received 13 abstracts, from which 8 articles were accepted.

Three articles present a method, 4 articles are original research, and 1 is a conceptual analysis. Overall they cover three broad Research Topics often discussed in the KG research communities: ontologies design, data architectures, reasoning.

Ontologies are the back-bone of KG modeling as they define what is in the data and how the information is connected. The Research Topic covers this import topic with three publications:

- "*C3PO: a crop planning and production process ontology and knowledge graph*" by Darnala et al. presents the design method to build and update a modular ontology and associated knowledge graph about vegetable production and planification activities. Some new design patterns are defined dedicated to agriculture. For example, the set of planned tasks that compose a technical itinerary of a crop type are presented. The final C3PO knowledge graph was used by the Elzeard enterprise to build three decision information systems.

- "*EPPO ontology: a semantic-driven approach for plant and pest codes representation*" by Ayllón-Benitez et al. presents the translation of European and Mediterranean Plant Protection Organization (EPPO) database into an OWL ontology. Each entity identified by an EPPO code becomes an OWL class. The ontology will be used as lingua franca to search data into different information systems used in BASF.
- "*Ontological how and why: action and objective of planned processes in the food domain*" by Dooley and Naravane present an extension of the FoodOn ontology about food processes. They propose two new types of process representations: processes by objectives, processes by mechanisms. Their goals are to improve search capability and identification.

An ontology on its own is not much use without data to instantiate it. The past decades of research into KG saw several approaches being presented to combine and align different data into a KG. Not all of those apply straight away to the agricultural domain and this Research Topic features 4 articles proposing specialized innovative approaches:

- "*CowMesh: a data-mesh architecture to unify dairy industry data for prediction and monitoring*" by Pakrashi et al. presents an approach to integrate data in the dairy industry by leveraging a combination of data mesh and data fabric design pattern. The approach is presented from a general point of view along with two specific use-case examples for the dairy industry.
- "*Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case study on grain legumes*" by Imbert et al. presents the design method of a Neo4J graph database that integrates the trait and gene information extracted from several sources. The graph model reuses existing ontologies like the Gene Ontology (GO), the Plant Ontology (PO) and the Plant Experimental Condition Ontology (PECO). The method was applied on the database design related to five legume species.
- "*Combining different points of view on plant descriptions: mapping agricultural plant roles and biological taxa*" by Amardeilh et al. presents some guidelines to publish a mapping dataset between two knowledge graphs: The French Crop Usage thesaurus defined crop usage expressed in French. TAXREF is the nomenclatural and taxonomic repository of living organisms that appear in French territories. A new specialized RDF vocabulary of mapping is defined and presented.
- "*Integrating collective know-how for multicriteria decision support in agrifood chains—application to cheesemaking*" by Buche et al. presents a multi-criteria decision support system (MDCSS) based on the capture and modelization of collective know-how in a Knowledge Graph. The ontology for expressing this information is introduced together with an example application for the process of cheese making.

Lastly, to illustrate the "Knowledge" part of a KG and reasoning over this knowledge, we have in this issue one paper covering using a KG to infer new information:

- "*Using knowledge graphs to infer gene expression in plants*" by Thessen et al. illustrates how a knowledge graph connecting partial information available about different plants can lead to new insights. Leveraging homologous genes as an inference back-end it is possible, as shown, to infer some of the unknown phenotypic impacts of plants gene regulatory networks.

We would like to thank the authors who submitted articles, the reviewers who evaluated them and the external editors who managed the reviews. All these people helped build a quality program for this Research Topic.

Author contributions

CR: Writing—original draft, Writing—review & editing. CG: Writing—original draft, Writing—review & editing. M-AL: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. CR benefits from support of the Data to Knowledge in Agronomy and Biodiversity (D2KAB-www.d2kab.org) project that received funding from the French National Research Agency (ANR-18-CE23-0017) and the project Partages de Connaissances (PACON) of the transverse program MetaBio funded by INRAE.

Acknowledgments

CR was part of the French association for Artificial Intelligence (AFIA https://afia.asso.fr/) and would like to thank the knowledge engineering college of AFIA for their help in the reviewing process.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 28 April 2023
DOI 10.3389/frai.2023.1145007

Integrating collective know-how for multicriteria decision support in agrifood chains—application to cheesemaking

Patrice Buche[1*†], Julien Couteaux[2†], Julien Cufi[1†], Sébastien Destercke[3†] and Alrick Oudot[1†]

[1]IATE, INRAE, Univ. Montpellier, Institut Agro, Montpellier, France, [2]I2M, INRAE, Univ. Bordeaux, Bordeaux, France, [3]HEUDIASYC, CNRS, Univ. Compiègne, Compiègne, France

Agrifood chain processes are based on a multitude of knowledge, know-how and experiences forged over time. This collective expertise must be shared to improve food quality. Here we test the hypothesis that it is possible to design and implement a comprehensive methodology to create a knowledge base integrating collective expertise, while also using it to recommend technical actions required to improve food quality. The method used to test this hypothesis consists firstly in listing the functional specifications that were defined in collaboration with several partners (technical centers, vocational training schools, producers) over the course of several projects carried out in recent years. Secondly, we propose an innovative core ontology that utilizes the international languages of the Semantic Web to effectively represent knowledge in the form of decision trees. These decision trees will depict potential causal relationships between situations of interest and provide recommendations for managing them through technological actions, as well as a collective assessment of the efficiency of those actions. We show how mind map files created using mind-mapping tools are automatically translated into an RDF knowledge base using the core ontological model. Thirdly, a model to aggregate individual assessments provided by technicians and associated with technical action recommendations is proposed and evaluated. Finally, a multicriteria decision-support system (MCDSS) using the knowledge base is presented. It consists of an explanatory view allowing navigation in a decision tree and an action view for multicriteria filtering and possible side effect identification. The different types of MCDSS-delivered answers to a query expressed in the action view are explained. The MCDSS graphical user interface is presented through a real-use case. Experimental assessments have been performed and confirm that tested hypothesis is relevant.

KEYWORDS

ontology, decision support systems, Semantic Web, knowledge representation, expertise integration, cheesemaking

OPEN ACCESS

EDITED BY
Christophe Guéret,
Accenture, Ireland

REVIEWED BY
Donagh Berry,
Teagasc Food Research Centre, Ireland
Stephanie Bisgaard,
Accenture, United States
Jan Top,
Wageningen University and
Research, Netherlands

*CORRESPONDENCE
Patrice Buche
✉ patrice.buche@inrae.fr

†These authors have contributed equally to this work

RECEIVED 15 January 2023
ACCEPTED 03 April 2023
PUBLISHED 28 April 2023

CITATION
Buche P, Couteaux J, Cufi J, Destercke S and Oudot A (2023) Integrating collective know-how for multicriteria decision support in agrifood chains—application to cheesemaking. *Front. Artif. Intell.* 6:1145007.
doi: 10.3389/frai.2023.1145007

1. Introduction

Agrifood chain processes are based on a multitude of knowledge, know-how and experiences forged over time. Agrifood companies that manage food product processing rely on their know-how to tailor their practices to the prevailing raw material variations, consumer expectations and regulations. The practice of acquiring knowledge through hands-on experience is a common one in the transformer industry, resulting in a vast accumulation of expertise among workers. This knowledge is typically passed on through on-the-job training and learning by doing. However, recent economic and health crises, along with internal changes within companies such as increased turnover and difficulty recruiting in certain sectors, have made it increasingly challenging to preserve and transmit this valuable know-how.

The aim of this paper, building upon the work of Buche et al. (2019), is to develop a new method for gathering and organizing knowledge, integrated in a software tool that can aid in preserving, accessing, and regularly updating the collective knowledge of the food industry for use in technology-related decision making. By implementing this methodology, we hope to overcome the challenges faced in preserving and transmitting the wealth of expertise within the industry and support the continued development of the food sector. The possibility of sustainably safeguarding and promoting practitioners' experience, as well as the technical expertise and scientific knowledge gained within a given food processing chain will be demonstrated based on a long-term collaboration with French cheesemaking companies with a "geographical indication" label, such as the protected designation of origin [*appellation d'origine protégée* (AOP)] and protected geographical indication [*indication géographique protégée* (IGP)].

The emergence of methods based on knowledge engineering in the field of food and bio-based product processing facilitates the development of decision-support tools that model complex reasoning based on processing operators' expertise (Buche et al., 2019; Baudrit et al., 2022; Belaud et al., 2022; Munch et al., 2022). Here we present a new multicriteria decision-support system (MCDSS) based on collective know-how which enables the formulation of recommendations on technological actions that may help maintain product quality or correct a product quality defect at the scale of a given food processing operation.

The MCDSS workflow process presented in Figure 1 consists of five main steps. The first one is a collaborative mind mapping activity involving almost all technicians of a given food chain and coordinated by a technical expert serving as an adviser in each chain. He/she is responsible for structuring the knowledge expressed in decision trees using a mind mapping software tool that respects some simple syntactic conventions (keyword labels in nodes). One decision tree is associated with a situation of interest (a product quality or defect) while being input in a given mind-mapping file. A decision tree represents potential causal relations between the situation of interest and explanatory situations associated with recommendations in terms of technological actions to manage the situation of interest. The second step involves individually and then collectively determining the efficiency of actions based on technician feedback. This information is input in the same mind-mapping file. In the third step, the mind-mapping file is automatically translated and stored in the knowledge base implemented as an RDF knowledge graph. End-users (technicians, food chain operators, students, etc.) mine, in the fourth step, the knowledge base using two views available in the MCDSS to deliver recommendations. For a given situation of interest, the explanatory view displays all possible explanatory situations, associated analytical parameter values and technical actions to correct/reach the situation of interest. The Action View feature enables users to efficiently filter actions based on multiple criteria within a decision tree, in order to correct or reach a desired situation. Additionally, it allows users to identify any potential side effects associated with a given recommendation. Users can easily switch back and forth between the two views, facilitating the process of selecting the best recommendation for a specific situation. The MCDSS workflow process is iterative (see fifth step in Figure 1), i.e., each decision tree including action efficiency indicators may be easily updated in the mind mapping tool to account for new experiences which are then automatically translated in the MCDSS knowledge base.

The Materials and methods section focuses on the following topics:

- Specifications and architecture of the decision-support system.
- A proposed model to aggregate individual action efficiency assessments.
- An ontological model to structure MCDSS knowledge base content.
- Two views of the multicriteria decision-support system.

The Results and Discussion section presents MCDSS functionality assessments and a comparison with the current state of the art.

2. Materials and methods

2.1. Specifications and architecture of the decision-support system

The detailed MCDSS specifications were determined in collaboration with several technical centers associated with French cheesemaking, namely Comté, Reblochon, Emmental de Savoie, Cantal, and Salers in the framework of two research projects funded by the French government from 2017 to 2023 (CASDAR Docamex, France Relance Docamex). Hereafter is a list of target functionalities:

1. For a given situation of interest (targeted food quality or defect), the MCDSS must provide all known possible explanations organized in a decision tree starting from the most general explanatory situations, which must be refined by more specific explanations until it is precise enough to propose an action lever and an associated recommended technological action. It must represent interactions between explanatory situations. Two kinds of interaction should be considered: (i) conjunctive interactions of situations S_1 and S_2 to explain S_3, which means that situation S_3 may emerge only if S_1 and S_2 appear; (ii) strengthening (resp. weakening) interactions of situation S_1 by situation S_2 to explain S_3, which means that the effect of S_1 on S_3 is strengthened (resp. weakened) if S_2 appears. The decision tree will enable users to consider all possible known explanations of a given situation of interest. This functionality, which is mostly geared toward junior technicians, is very important in cheesemaking chains as they have to deal with growing turnover rates.
2. It should be possible to associate a situation ("of interest" or explanatory) with the value of a relevant analytical parameter that allows verification that the situation is actually happening.

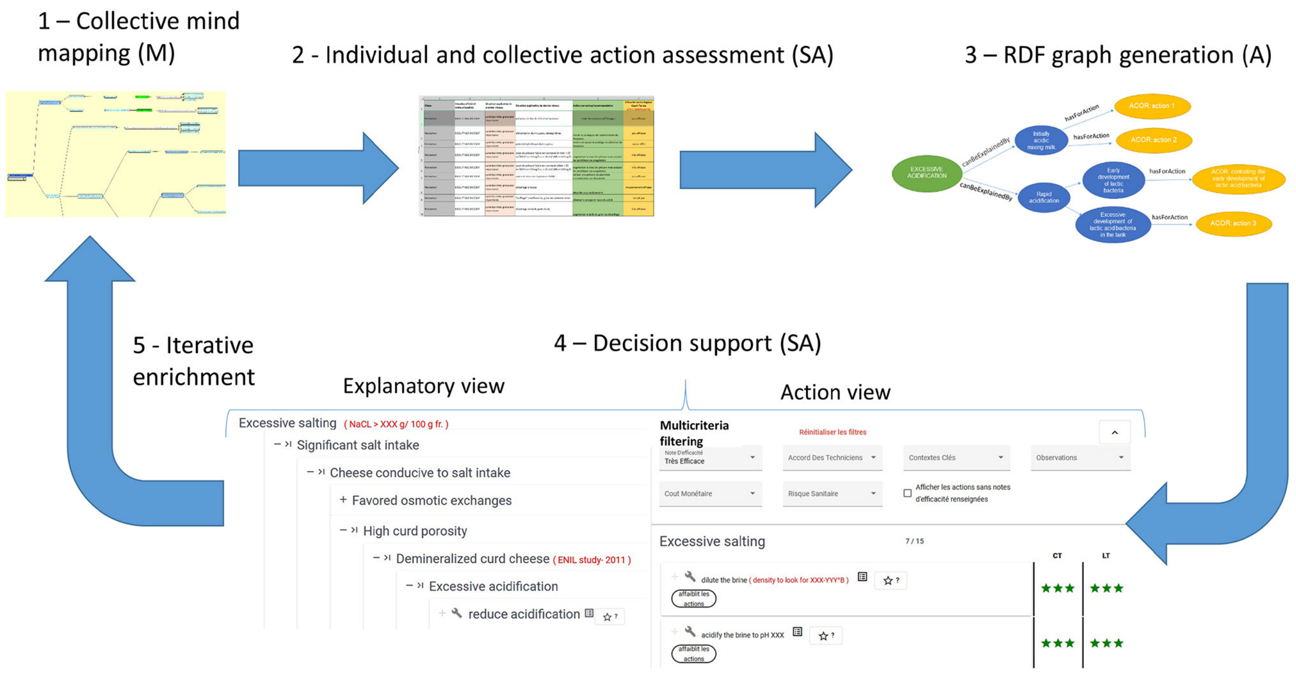

1 – Collective mind mapping (M)
2 - Individual and collective action assessment (SA)
3 – RDF graph generation (A)
EXCESSIVE ACIDIFICATION
canBeExplainedBy
canBeExplainedBy
Initially acidic mixing milk
Rapid acidification
Early development of lactic bacteria
Excessive development of lactic acid bacteria in the tank
hasForAction
hasForAction
hasForAction
hasForAction
ACOR: action 1
ACOR: action 2
ACOR: controlling the early development of lactic acid bacteria
ACOR: action 3
5 - Iterative enrichment
4 – Decision support (SA)
Explanatory view
Action view
Excessive salting (NaCL > XXX g/ 100 g fr.)
Significant salt intake
Cheese conducive to salt intake
Favored osmotic exchanges
High curd porosity
Demineralized curd cheese (ENIL study- 2011)
Excessive acidification
reduce acidification
Multicriteria filtering
Réinitialiser les filtres
Note D'efficacité
Très Efficace
Accord Des Techniciens
Contextes Clés
Observations
Cout Monétaire
Risque Sanitaire
Afficher les actions sans notes d'efficacité renseignées
Excessive salting
7 / 15
CT
LT
dilute the brine (density to look for XXX-YYY°B)
affaiblit les actions
★★★
★★★
acidify the brine to pH XXX
affaiblit les actions
★★★
★★★

FIGURE 1
Workflow process associated with the MCDSS (M, manual task; A, automatic task; SA, semi-automatic task). The three stars indicate the action's efficiency is "very effective".

This is of great interest for technicians who have to deal with several cheese production processes (e.g., Comté and Bleu de Bresse for the CTFC technical center) without being fully aware of all of the analytical parameter values associated with the encountered situations.

3. The MCDSS must be able to determine the possible side effects of an action: a corrective action for one situation of interest should not lead another problem.
4. Feedback on technicians' individual experiences in terms of technological action efficiency to deal with a given situation of interest must be registered and aggregated. Indeed, action ranking is of great importance to help users choose the "best" action to cope with a given situation of interest. Moreover, registration of contextual criteria relevant for decision-support and associated with those assessments is required to facilitate decision support. For instance, a given action like "Review herd rationing practices" may be considered very efficient in the long term (LT), yet not at all efficient in the short term (ST). The MCDSS must be able to rank actions using a multicriteria filtering system.
5. It must represent the expert knowledge expressed in decision trees using international World Wide Web Consortium (W3C) standards in order to facilitate interoperability between industry and academic institutes in an Open Science setting. More particularly, two standard languages are recommended: (1) Resource Description Framework (RDF) for graph data description and exchange. RDF provides a variety of syntax notations and data serialization formats; (2) Web Ontology Language (OWL), a family of knowledge representation languages for authoring RDF-based ontologies.

2.2. From mind mapping to formal knowledge representation

Buche et al. (2019) proposed a method that enables collective mind mapping dedicated to this MCDSS. Interested readers may refer to this paper for further details on step 1 implementation (see Figure 1). In this section, we focus on two new contributions of the paper. The first concerns a numerical model that aggregates individual assessments associated with action efficiency expressed by technicians into a single indicator. This functionality is required in Specification 4 (see Section 2.1). During step 2, as presented in Figure 1, the aggregated indicator is discussed and validated collectively by the team of technicians to determine the final action efficiency value, which is input in the knowledge base for decision-making support. The second contribution is an extended version of the ontology presented in Buche et al. (2019) to structure the information in the MCDSS knowledge base for navigation and querying purposes. The extension includes the efficiency indicators and associated criteria. This extended version is expressed using the W3C standards to fulfill Specification 5 (see Section 2.1), which is also a novel contribution of this paper as the ontology presented in Buche et al. (2019) was based on the Conceptual Graph model (Sowa, 1984; Chein and Mugnier, 2009).

2.2.1. A model to aggregate individual action efficiency assessments

Each technician, denoted T_i hereafter, provides two types of information:

- His/her experience in terms of number of action implementations, called F_i, reflecting the reliability of his/her statements, which takes its value in the set {(*N*)ever, (*R*)arely: 1 < 3, (*S*)ometimes: 3 < 10, (*O*)ften): >10}, as summarized by $R = \{N, R, S, O\}$.
- The efficiency of the action, denoted E_i, which takes its value in {Very effective (*A*), Moderately effective (*B*), Not very effective (*C*), No effect (*D*)}, as summarized by $E = \{A, B, C, D\}$.

The technician can also select "don't know" for the second value.

With the experience of the technician corresponding to the number of times (roughly) where he/she encountered the situation of interest, it seems quite natural to interpret his/her answer as a number of "virtual" observations. We will therefore associate with each value in *R* an equivalent number, i.e., *N*→ *0.5, R*→ *2, S*→ *5, O*→ *10.* In practice, each of these values is chosen to be within the corresponding interval. For instance, *Rarely* corresponds to the interval [1,3], for which we picked the central value 2. We still assigned a positive value to *Never*, so as to reflect the fact that the reported experience may come from sources other than direct observation. Those choices were made in accordance with the end user and can in practice be changed according to the application, as they remain subjective (but not arbitrary) to some extent. The corresponding intervals could in principle also be kept, yet processing such information would increase the cognitive load for users, hence our choice to keep precise numbers representing the numbers of experiments.

Let n_i be the number corresponding to the experience of technician T_i. For example, if technician T_i answers $F_i = R$, therefore rarely, then $n_i = 2$. If k technicians provide an answer, then *total* $N = \sum_{i=1}^{k} n_k$ will denote the total number of virtual observations.

The aim is then—based on these virtual observations—to construct a histogram on *E*, and associate a probability with each of its elements. Let n^A, n^B, n^C, n^D denote the total number of observations given to *A, B, C, D*, respectively.

Definition 1: n^A the total number of observations given to *A* is defined by

$$n^A = \sum_{T_i:\ E_i = A} n_i$$

The probability (subjective and *a priori*) of *A* then becomes

$$p(A) = \frac{n^A}{Nb}$$

and the same for *B, C, D*.

Example 1: Suppose three technicians provide their opinions as follows:

- $F_1 = R \Rightarrow\ n_1 = 2;\ E_1 = A$ (very effective)
- $F_2 = S \Rightarrow\ n_2 = 5;\ E_2 = B$ (moderately effective)
- $F_3 = R \Rightarrow\ n_3 = 2;\ E_3 = C$ (not very effective)

which gives $N = 9$ and *p(A)* = 2/9; *p(B)* = 5/9; *p(C)* = 2/9.

The information given by the previous distribution is probably too complex to be readily understood by a technician and requires a simple summary. This can easily be done through various statistics and then supplied to the user in graphical and easily interpretable form. In contrast with number of times a situation has been encountered, in our case efficiency is not associated with an actual numerical measure. Moreover, such measures would probably vary across situations and not be comparable. We therefore chose to not replace ordered categories A, B, C by numbers, and instead provided both a central value and its dispersion based on the quantile notion. More precisely, we will use the median (quantile at 50%) and two quantiles around the latter (therefore 50% – α and 50% + α) as a statistical summary.

Definition 2: the quantile of level ∈ [0, 1], denoted i_β, relative to the distribution *p* defined on *E* is the value

$$i_\beta = \left\{ j \in E :\ \left(\sum_{l \le j-1} p(l) < \beta \right) \bigwedge \left(\sum_{l \le j} p(l) \ge \beta \right) \right\}$$

where < corresponds to the alphabetical order and with the convention $\sum_{l \le 0} p\left(l\right) = 0$.

Let us get back to our previous example, where we will conventionally denote $P\left(\{A, B\}\right) = p\left(A\right) + p(B)$, etc.

Example 2-1:

P({A}) = p(A) = 2/9 = 0.2222.

P({A,B}) = p(A) + p(B) = 2/9 + 5/9 = 0.77777.

P{A,B,C}) = P({A,B,C,D}) = 1.

We will therefore have the following quantile $i_{0.1} = A$ (first decile) because $\sum_{l \le 0} p\left(l\right) = 0$ and $\sum_{l \le A} p\left(l\right) = 0.2222$ therefore$\left\{ \left(\sum_{l \le 0} p\left(l\right) < 0.1\right) \bigwedge \left(\sum_{l \le A} p\left(l\right) \ge 0.1\right) \right\}$ is true.

In the same way, $i_{0.25} = B$ (first quartile); $i_{0.5} = B$ (median); $i_{0.75} = B$ (third quartile); $i_{0.9} = C$ (ninth decile).

It is clear that if the technicians all provide the same evaluation, then all the quantiles will have the value of this evaluation. Conversely, if the technicians are somewhat divided and of equivalent experience, the difference between the quantiles will show this uncertainty. We hence propose to match A, B, C, D to a number of "stars" (3,2,1,0) and to provide the average of the values observed in set $[i_{0.1}, i_{0.9}]$ *as a reference value*. In our example, this is the set [A, B, C], with the reference value 2. It would also be useful to show that there is no consensus on this reference value by highlighting all the intervals [1,3].

Figure 7 presents examples of graphical representations in terms of stars. The following example illustrates the case where the reference value is not one of the initial values.

Example 2-2:

Suppose that two technicians provide their opinions as follows:

- $F_1 = R \Rightarrow\ n_1 = 2;\ E_1 = A$ (very effective)
- $F_2 = P \Rightarrow\ n_2 = 5;\ E_2 = B$ (moderately effective)

In this case, p(A) = 2/7, p(B) = 5/7 with $i_{0.5} = B$ (median) and $[i_{0.1}, i_{0.9}] = [A, B]$ with 2.5 being the obtained average (stars).

2.2.2. A new ontological model to structure the MCDSS knowledge base content

Decision trees edited in mind-map files in step 1 and enriched with action efficiency assessments in step 2 must be stored in the

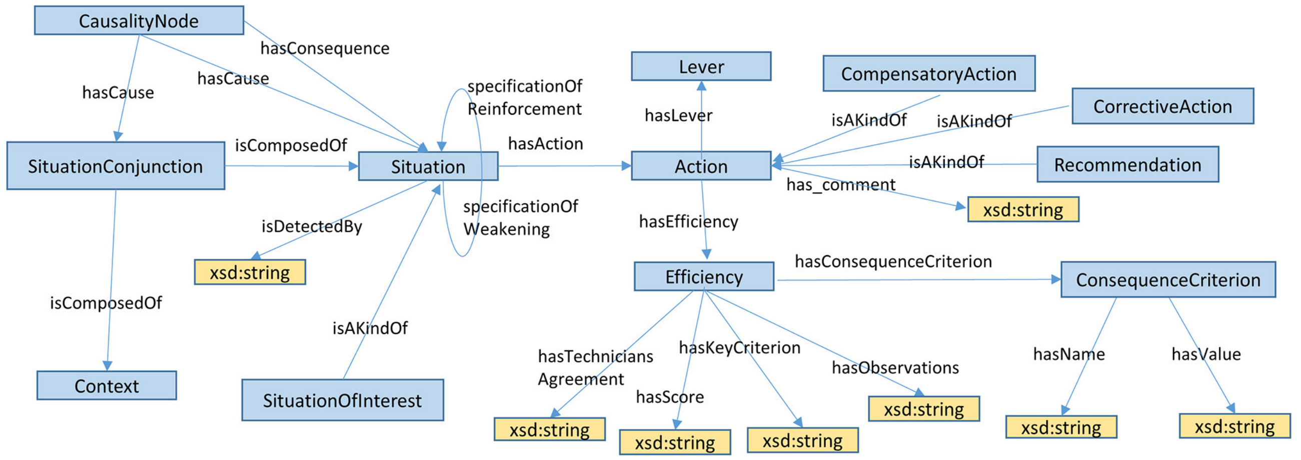


FIGURE 2
OWL ontological model used to structure a decision tree in the MCDSS knowledge base.

MCDSS knowledge base. As indicated in Specification 5, Semantic Web language standards created by W3C must be used for knowledge base implementation. The OWL ontology—an original contribution of this paper—designed to structure and instantiate a decision tree in the RDF knowledge base is presented in this section.

The OWL definition of classes and properties presented in Figure 2 is available in Buche et al. (2022). Hereafter we explain how this ontological model takes the specifications expressed in Section 2.1 into account.

As expressed in Specification 1, a situation S_1, an instance of the *Situation* OWL class, is explained by a situation S_2 through an instance of the *CausalityNode* class linked to S_1 (resp. S_2) by the OWL *hasForCause* (resp. *hasForConsequence*) object property. Note that S_1 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 CI_1, an instance of the *SituationConjunction* class, is linked to conjunctive causal situations S_1 and S_2 (and other situations if required) by the *isComposedOf* property. CI_1 is linked to an instance of the *CausalityNode* class by the *hasForCause* property. This CausalityNode class instance is linked to the consequence situation S_3 by the *hasForConsequence* property. The strengthening (resp. weakening) interaction of situation S_1 by situation S_2 to explain S_3 is also represented using a conjunctive interaction CI_1, an instance of the *SituationConjunction* class. The asymmetric role of situations is achieved in the following way: the altered situation S_1 is linked to altering situation S_2 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 Specification 2. An instance of *Action* is associated with an instance of the *Efficiency* class to implement Specification 4. The *hasForKeyCriterion* datatype property permits determination of the list of criteria values associated with a single *Efficiency* instance. The *hasForScore*, *hasForObservations*, and *hasForTechniciansAgreement* datatype properties are associated with an *Efficiency* class instance which 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.

Figure 3 is an excerpt of a mind-mapping file representing the decision tree associated with the situation of interest *Excessive salting* achieved by the blue node at the bottom left part of the figure. The entire mind-mapping file is available in Buche et al. (2022). This situation of interest may be explained by the *Significant salt intake* situation. Then four explanations are possible. Hereafter we will consider the one whose node is white, i.e., *Conditions favoring salt uptake in brine* and its associated branch, whose nodes are also white, until reaching the two nodes *Put the brine tank in the dryer* and EFFICIENCY: ST. Figure 4 shows a zoom on the table associated with the node EFFICIENCY: ST. This table includes the aggregated efficiency indicator with the number of observations (see Section 2.2.1) and contextual criteria associated with them. In Figure 5, we present a part of the MCDSS RDF knowledge base corresponding to the translation of the branch whose nodes are white in the decision tree presented in Figure 3. The entire RDF graph corresponding to the mind-mapping file is available in Buche et al. (2022).

In Figure 5, to facilitate the understanding of the translation of Figures 3, 4 into RDF, instances of OWL classes are represented by rectangles, with the class name in the header complemented by a pseudo-label representing its URI (as the real one is too long) or the associated value of the rdfs:label property. Values associated with datatype properties are framed in black.

2.3. Multicriteria decision-support system

The decision-support system (see step 4 in Figure 1) consists of two complementary access modes to the knowledge base content, i.e., the explanatory and action views. The explanatory view

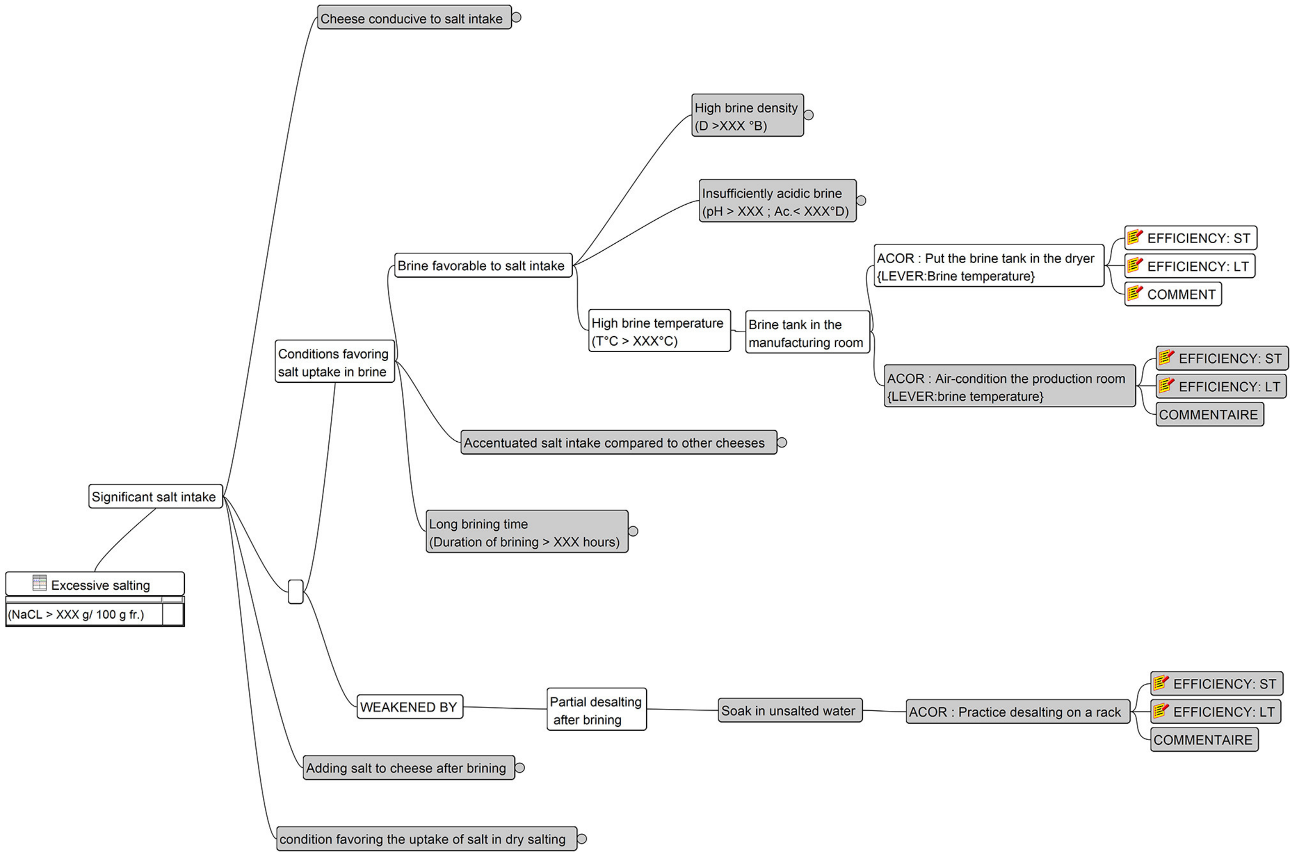
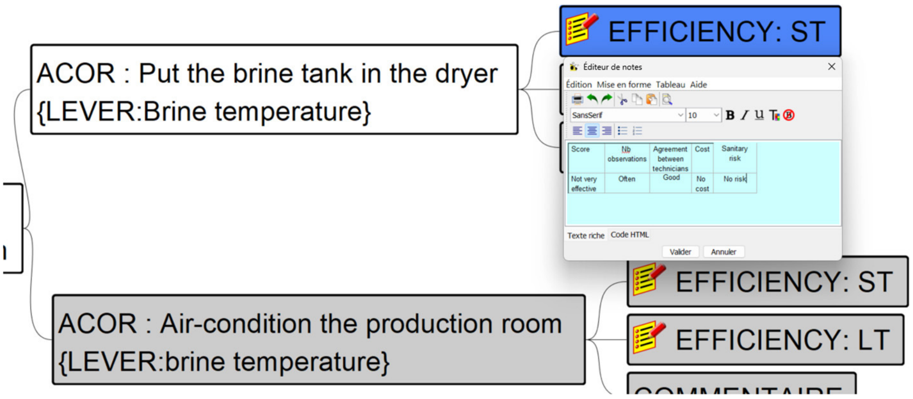

Cheese conducive to salt intake
High brine density
(D >XXX °B)
Insufficiently acidic brine
(pH > XXX ; Ac.< XXX°D)
Brine favorable to salt intake
ACOR : Put the brine tank in the dryer
{LEVER:Brine temperature}
EFFICIENCY: ST
EFFICIENCY: LT
COMMENT
High brine temperature
(T°C > XXX°C)
Brine tank in the
manufacturing room
Conditions favoring
salt uptake in brine
ACOR : Air-condition the production room
{LEVER:brine temperature}
EFFICIENCY: ST
EFFICIENCY: LT
COMMENTAIRE
Accentuated salt intake compared to other cheeses
Significant salt intake
Long brining time
(Duration of brining > XXX hours)
Excessive salting
(NaCL > XXX g/ 100 g fr.)
WEAKENED BY
Partial desalting
after brining
Soak in unsalted water
ACOR : Practice desalting on a rack
EFFICIENCY: ST
EFFICIENCY: LT
COMMENTAIRE
Adding salt to cheese after brining
condition favoring the uptake of salt in dry salting

FIGURE 3
An excerpt of the mind-mapping file associated with the *Excessive salting* situation of interest.


ACOR : Put the brine tank in the dryer
{LEVER:Brine temperature}
EFFICIENCY: ST
Éditeur de notes
Édition Mise en forme Tableau Aide
SansSerif
10
B I U
Texte riche
Code HTML
Valider
Annuler
EFFICIENCY: ST
EFFICIENCY: LT
COMMENTAIRE
ACOR : Air-condition the production room
{LEVER:brine temperature}

Score	Nb observations	Agreement between technicians	Cost	Sanitary risk
Not very effective	Often	Good	No cost	No risk

FIGURE 4
Zoom on the table associated with the EFFICIENCY: ST node present in the mind-mapping file associated with the *Put the brine tank in the dryer* node.

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/reach the situation of interest. The action view displays the list of actions related to a given decision tree to correct/reach the associated situation of interest. It enables multicriteria filtering, action ranking and side effect identification.

Both views may be used independently and jointly depending on the usage case. For instance:

- A systematic review of all possible explanatory situations is carried out using the explanatory view.
- Solving a contextualized problem is carried out using the action view through the multicriteria

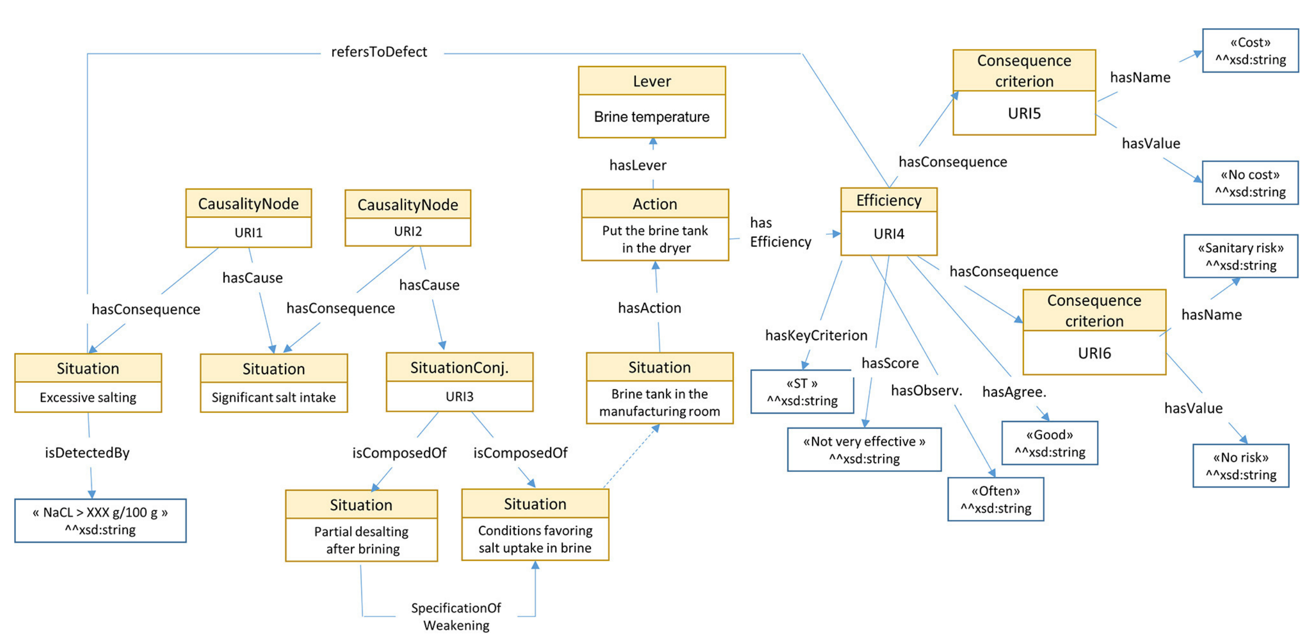


FIGURE 5
An excerpt of the MCDSS knowledge base corresponding to the selected branch in Figure 3.

filtering mode, sometimes complemented with the explanatory view.

Hereafter we define a multicriteria (MCDSS) query executed in the action view and the associated answers. Then we present the MCDSS graphical user interface (GUI) using an illustrative example based on a real case from a French protected designation of origin (AOP) chain.

2.3.1. MCDSS query definition

We define in this section the notion of MCDSS query *Q* executed on the *KB* knowledge base. Then the answer to *Q*, called *AN*, and the two complementary *AN-inter* and *AN-intra* answers are defined for side effect identification.

Definition 3: The MCDSS knowledge base (*KB*) is defined as the 9-tuple (*S*, V_k, C_c, V_c, *A*, *L*, *E*, *Ag*, *O*), with:

- *S* = the set of instances of the *SituationOfInterest* class;
- V_k = the set of key criteria labels associated with the *hasForKeyCriterion* datatype property;
- C_c = the set of consequence criteria names associated with the *hasForName* datatype property;
- V_c = the set of consequence criteria values associated with the *hasForValue* datatype property;
- *A*= the set of *Action* class instances implemented using *Lever* class instances;
- *L*= the set of *Lever* class instances;
- *E*= the set of action efficiency labels associated with the *hasForScore* datatype property;
- *Ag*= the set of action efficiency consensus labels associated with the *hasForTechnicianAgreement* datatype property;
- *O*= the set of action efficiency labels associated with the *hasForObservations* datatype property.

Definition 4: Given *KB* defined in Def. 3, the set of input conjunctive filtering parameters associated with an MCDSS query *Q* executed in *KB* is defined by the 6-tuple:

$$(s \in S,\ \{v_1, \cdots, v_m\} \in V_k,\ \{(c_1, v_1), \cdots, (c_n, v_n)\} \epsilon (C_c, V_c)^n,\ \{e_1, \cdots, e_o\} \in E,\ \left\{ag_1, \cdots, ag_p\right\} \in Ag,\ \left\{o_1, \cdots, o_q\right\} \in O)$$

Note: multivalued parameters are considered to be aggregated disjunctively in the querying.

Example 3: *Q1* = (*excessive salting*,{∅},{∅}, $\{$*very effective*$\}$, $\{$*good, average*$\}$, {∅}) represents the querying of the *excessive salting* situation of interest with the action efficiency being *very effective* and the action efficiency consensus being *good* or *average*. The SPARQL query generated by the MCDSS and corresponding to *is* available in Buche et al. (2022).

Definition 5: The answer *AN* associated with an MCDSS query *Q* executed in *KB* is defined by a set of 2-tuples: $\left\{(a_1, l_1), \cdots, (a_n, l_n) \epsilon (A, L)^n\right\}$, with (a_i, l_i) related to the decision tree associated with the situation of interest *s*.

Example 4: *AN1* = {*(dilute the brine, Brine salt concentration), (acidify the brine to pH 5.4, Brine acidity), (practice brining on a rack, Brining equipment), (reduce brining time, Brine duration)*} is the answer that includes the four recommended actions associated with the *query* of Example 1. The triples results of the SPARQL query corresponding to *is* available in Buche et al. (2022).

Two complementary answers with *AN* are provided by the MCDSS when the *Q* query is executed. The objective, corresponding to Specification 3 (see Section 2.1), is to identify two types of potential side effects that could occur if a recommended action related to *AN* is implemented:

- *AN-inter*: potential side effects with other situations of interest related to *KB*. Situations where the associated decision tree recommends the use of a lever associated with a given *AN* action are selected.
- *AN-intra*: potential side effects with other actions related to the decision tree associated with the situation of interest expressed in *Q*.

Definition 6: Given the *AN* answer to a query Q, the *AN-inter* answer associated with a recommended a_i *action* implemented using a given l_i lever with $(a_i, l_i) \in AN$ is defined by a set of 2-tuples:

$(s' \in S, \{(a_1, l_i), \cdots, (a_n, l_i) \epsilon (A, L)^n\})$ with $s' \neq s$, with s being the situation of interest associated with the *Q* query and (a_j, l_i), *j=1, ...n* being related to the decision tree associated with the *s'* situation of interest.

Example 5: *AN-inter1* associated with the recommendation *(reduce brining time, Brine duration)* related to *AN1* is *{(unpleasant taste or odor*, {*(extend the brining time to 2 h maximum*, *Brine duration)}),(brown paste*,{*(extend the brining time to 2 h maximum*, *Brine duration)}), (excessive proteolysis*,{*(extend the brining time to 2 h maximum*, *Brine duration)}), (insufficient salting*, {*(extend the brining time to 2 h maximum*, *Brine duration)})}.*

AN-inter1 means that implementing the recommendation *reduce brining time* to solve the *excessive salting* situation may create a side effect with four other situations of interest likely to occur: *unpleasant taste or odor, brown paste, excessive proteolysis, insufficient salting.* Indeed, the same *Brine duration* lever is recommended to solve these situations but it is used in an opposite way *(extend the brining time to 2 h maximum)*, which could potentially trigger those situations of interest if the recommendation is applied. MCDSS users may query the decision trees associated with those situations of interest to find a good trade-off to avoid triggering unwanted side effects.

Definition 7: The *AN-intra* answer associated with an *a recommended action* implemented using a given *l* lever to solve the $s \in S$ situation of interest is defined by a 4-tuple:

$(\{(a_{11}, l_1), \cdots, (a_{1n}, l_n)\} \epsilon (A, L)^n\}, \{(a_{21}, l_1), \cdots, (a_{2n}, l_n)\} \epsilon (A, L)^n\}, \{(a_{31}, l_1), \cdots, (a_{3n}, l_n)\} \epsilon (A, L)^n\}, \{(a_{41}, l_1), \cdots, (a_{4n}, l_n)\} \epsilon (A, L)^n\})$ with a_{1i} actions (resp. a_{2i} actions) corresponding to potential weakening actions of the recommended *a* action (resp. potential actions weakened by the recommended *a* action) and a_{3i} actions (resp. a_{4i} actions) corresponding to potential reinforcement actions of the recommended *a* action (resp. potential actions reinforced by recommended *a* action).

Example 6: *AN-intra1* associated with the recommendation *(reduce brining time, Brine duration)* related to *AN1* is *({(practice desalting on a rack, Brining equipment)}),{∅},{∅}{∅}).*

AN-intra1 means that implementing the *reduce brining time* recommendation to solve the *excessive salting* situation may be weakened by the *practice desalting on a rack* action. Complementary information about this possible interaction may be found using the explanatory view.

2.3.2. MCDSS graphical user interface

Using an illustrative example, we show how the MCDSS graphical user interface has been implemented to propose both complementary access modes to the knowledge base content, i.e., the explanatory and action views.

The explanatory view proposes navigation in a decision tree associated with a given situation of interest to query all possible explanatory situations. Figure 6 shows an excerpt of the explanatory view for the *Excessive salting* situation of interest. Analytical values associated with situations are shown in red. For example, *NaCl rate* > *XXX* [1]*g/100g* is the value associated with the *Excessive salting* situation. The first high-level explanatory situation is *Significant salt intake by the cheese during its production*, while several others specify this high-level explanation. For instance, it could be explained by the *Conditions favoring salt uptake in brine* situation. By following this branch of the decision tree, we reach a more detailed explanation, i.e., *Too much salt added in brine*. This latter explanation is associated with the *Dilute the brine* action. Its associated analytical value is *Density to reach XXX-YYY°B*.

The action view enables knowledge base querying and filtering to solve a contextualized problem. Figure 7 shows a query presented in example 3 concerning the *Excessive salting* situation of interest. Filtering criteria used regarding the action efficiency indicator and agreement level enable filtering of three actions out of a total of 15 present in the decision tree.

Complementary answers identifying possible side effects may be obtained using buttons (see the two buttons at the bottom of Figure 7 corresponding to AN-intra and AN-inter answers for the *Reduce brining time* action). Figure 8 shows a list of four situations of interest presented in Example 5 above: *unpleasant taste or odor, brown paste, excessive proteolysis*, and *insufficient salting.* The *Brine duration* lever is recommended to solve these situations, while using it in an opposite way *(extend the brining time to 2 h maximum)* compared to that recommended for the *Excessive salting* situation of interest. Figure 9 shows the action presented in Example 5, which may weaken the recommended *Reduce brining time* action.

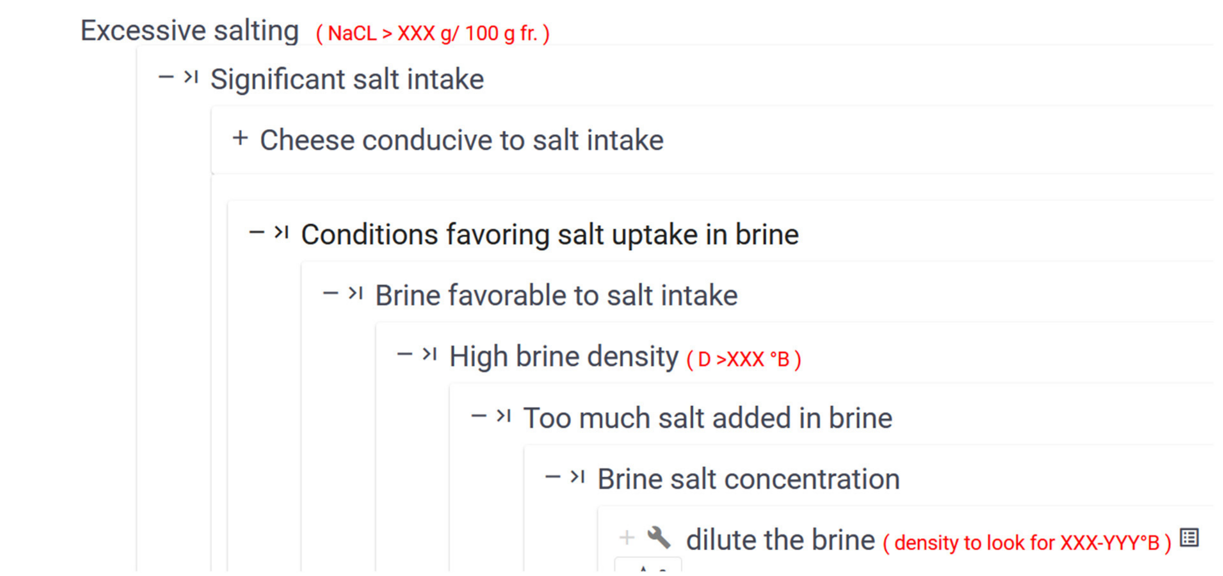
3. Results

In this section, we present the assessment results of Specifications 1, 2, and 4, which were performed with end-users.

3.1. Reviewing all possible technological actions associated with a situation of interest (Specifications 1 and 2)

In a technological reasoning task, for a given situation of interest (targeted quality or defect), a technician must be able to check all possible explanatory situations and corresponding analytical parameters to check that this situation will happen. Moreover, he/she must be aware of the associated recommended technological action. The protocol presented in Figure 10 was

[1] The actual numerical values have been anonymized to avoid recognition of the cheese chain.



Excessive salting (NaCL > XXX g/ 100 g fr.)

− ›I Significant salt intake

+ Cheese conducive to salt intake

− ›I Conditions favoring salt uptake in brine

− ›I Brine favorable to salt intake

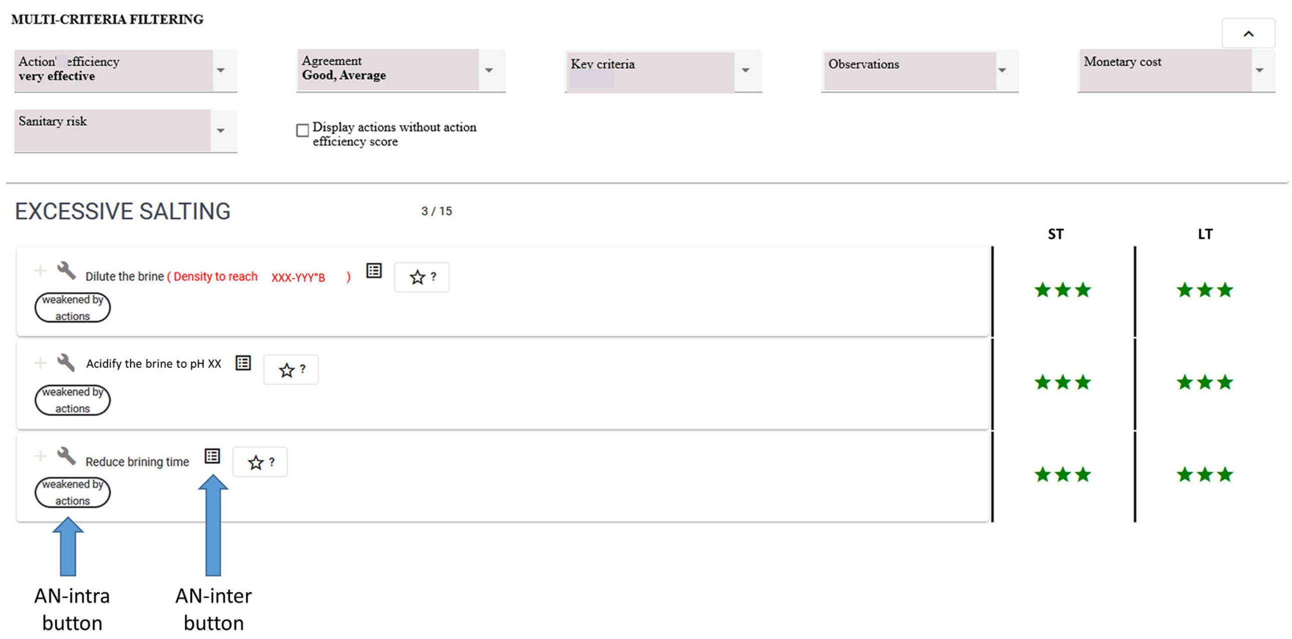
− ›I High brine density (D >XXX °B)

− ›I Too much salt added in brine

− ›I Brine salt concentration

+ dilute the brine (density to look for XXX-YYY°B)

FIGURE 6
MCDSS explanatory view showing an excerpt of the decision tree associated with the *Excessive salting* situation of interest.



MULTI-CRITERIA FILTERING

Action' efficiency: very effective
Agreement: Good, Average
Key criteria
Observations
Monetary cost
Sanitary risk
□ Display actions without action efficiency score

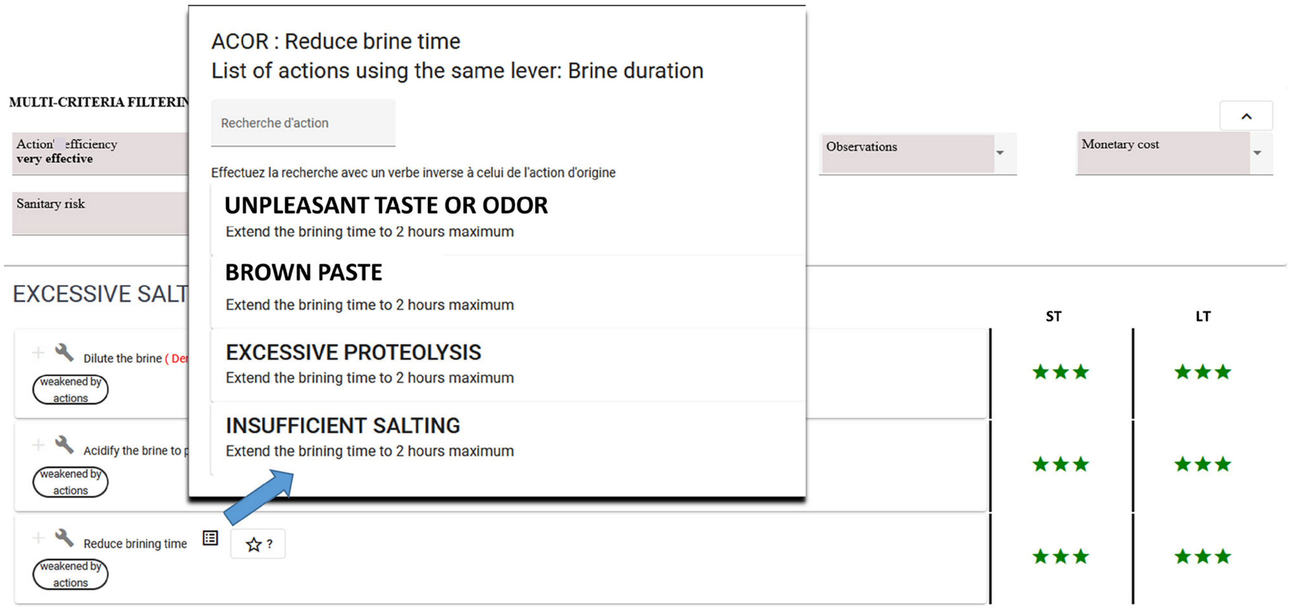
EXCESSIVE SALTING 3 / 15

	ST	LT
Dilute the brine (Density to reach XXX-YYY°B) ☆ ? weakened by actions	★★★	★★★
Acidify the brine to pH XX ☆ ? weakened by actions	★★★	★★★
Reduce brining time ☆ ? weakened by actions	★★★	★★★

FIGURE 7
MCDSS action view showing an excerpt of the list of filtered actions related to the decision tree associated with the *Excessive salting* situation of interest. The three stars indicate the action's efficiency is "very effective".

designed to assess the impact of MCDSS use in this reasoning task. It was tested with technicians from three different food chains. The protocol includes the following steps:

- Fifteen people related to three different chains passed this test. Figure 11 shows that 39% of the technological actions were noted without the MCDSS and 66.5% after its use, which represents 27% enhancement. Only two chains (10 people) undertook the analytical value tests as chain 1 corresponds to generic knowledge associated with a situation of interest learned in technical schools. As analytical values highly depend on a given cheesemaking process, it was not possible to conduct this test on this generic knowledge. Figure 12 shows that 18.33% of the correct answers were obtained without the MCDSS. The score increased to 76.25% after its use, which represents 60% enhancement. Both tests showed a good (even very good for the second one) enhancement with regard to the answers provided via use of the MCDSS.



ACOR : Reduce brine time
List of actions using the same lever: Brine duration

Recherche d'action

Effectuez la recherche avec un verbe inverse à celui de l'action d'origine

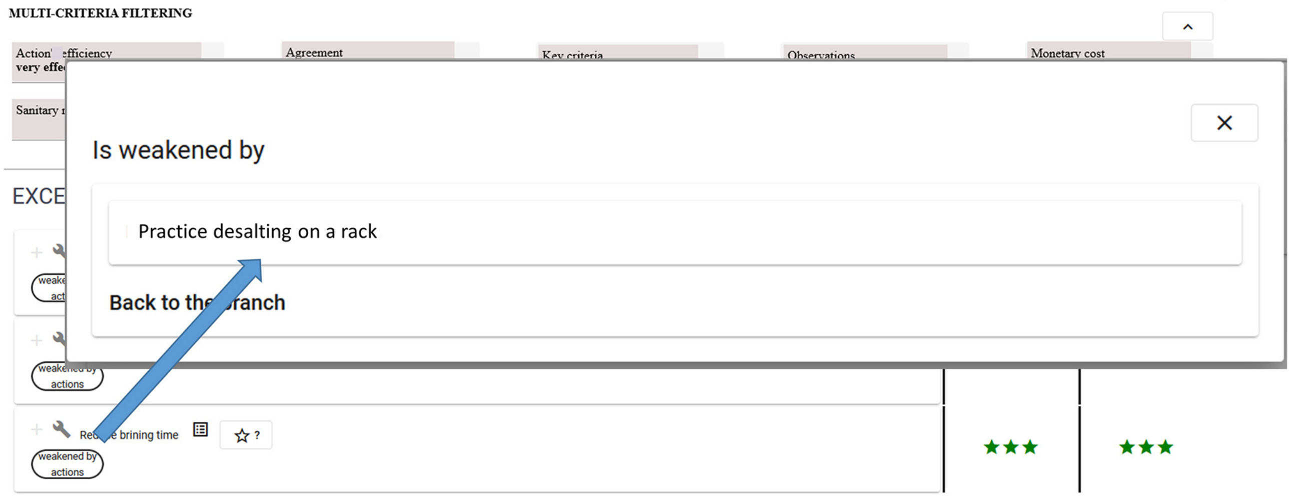
UNPLEASANT TASTE OR ODOR
Extend the brining time to 2 hours maximum

BROWN PASTE
Extend the brining time to 2 hours maximum

EXCESSIVE PROTEOLYSIS
Extend the brining time to 2 hours maximum

INSUFFICIENT SALTING
Extend the brining time to 2 hours maximum

FIGURE 8
MCDSS action view showing a list of four situations of interest using the same lever but in an opposite way compared to that recommended to solve the *Excessive salting* situation of interest. The three stars indicate the action's efficiency is "very effective".



Is weakened by

Practice desalting on a rack

Back to the branch

FIGURE 9
MCDSS action view showing a list of actions related to the decision tree associated with the *Excessive salting* situation of interest which could potentially weaken the recommended action. The three stars indicate the action's efficiency is "very effective".

- Unfortunately, the MCDSS prototype was not finished when the assessment campaign was carried out during the project. Consequently, it was not possible to assess the implementation corresponding to Specification 3 (identification of side effects associated with a recommended action). This will be of course done as soon as possible in the future. Nevertheless, the assessment results presented above suggest that these results will be also good. Indeed, finding all side effects between situations of interest (see Section 2.3, ANS-inter) may be a huge manual task as more than a 100 decision trees may be defined for a given cheese chain.

3.2. Technological efficiency aggregation assessment (Specification 4)

This assessment was conducted with a group of five technicians, all of whom were experts of a real cheesemaking process. It was focused on a set of three decision trees corresponding to three situations of interest (*Excessive dripping*, *Excessive acidification*, and *Excessive salting*). The five technicians provided individual assessments for 44 actions related to the three decision trees. Each action was assessed twice (88 assessments), with each assessment corresponding to two different production approaches: production

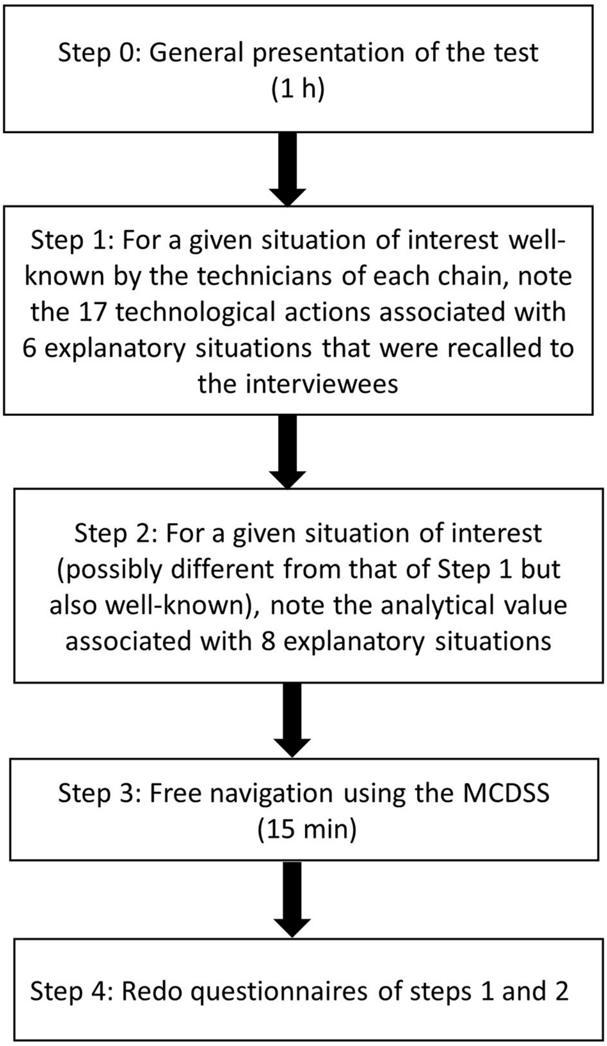

Step 0: General presentation of the test (1 h)
Step 1: For a given situation of interest well-known by the technicians of each chain, note the 17 technological actions associated with 6 explanatory situations that were recalled to the interviewees
Step 2: For a given situation of interest (possibly different from that of Step 1 but also well-known), note the analytical value associated with 8 explanatory situations
Step 3: Free navigation using the MCDSS (15 min)
Step 4: Redo questionnaires of steps 1 and 2

FIGURE 10
Protocol designed to assess the impact of MCDSS use in the reasoning task.

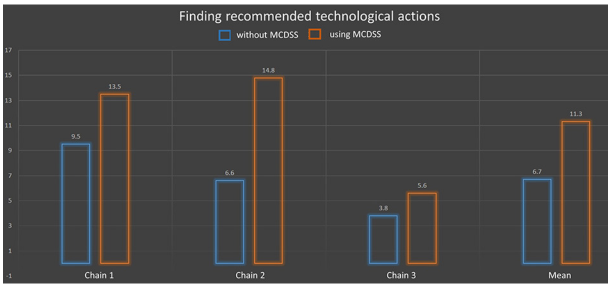


FIGURE 11
Test results (blue without MCDSS, brown with MCDSS) for recommended technological action findings: the y-axis represents the number of correct answers (mean value) in the three chains and on average.

approach 1 and production approach 2. Aggregated values were computed using the model presented in Section 2.2.1. In parallel, this group of experts collectively determined an assessment for each action without using the model. Computed assessments were compared to collective assessments. The associated data are available in Buche et al. (2022). Two assessments were considered to be in disagreement when there was a difference of at least two modalities (e.g., <no effect, moderately effective>, <very effective, not very effective>, etc.). The results presented in Table 1 show an error rate of around 5.7% (Total number of disagreements/Total number of actions), which is rather low. In practice, the method was considered relevant enough to compute an aggregated efficiency indicator associated with an action in a given decision tree. This aggregated indicator was discussed and validated collectively by the group of technicians during monthly meetings before being input in decision trees stored in the MCDSS knowledge base.

As already said above, the MCDSS prototype was not finished when the assessment campaign was carried out. Consequently, it was not possible to assess the implementation corresponding to Specification 4 about multicriteria filtering (identification of side effects associated with a recommended action).

4. Discussion

We discuss in this section the original contributions of the paper compared to the current state of the art, summarize and provide complementary information about key contributions and present some future directions of the research.

4.1. Comparison to the current state of the art

A lot of progress has been achieved in knowledge integration and multicriteria analysis methods and tools in the food science and technology field, yet they remain fragmented and incomplete (Aceves Lara et al., 2018; Thomopoulos et al., 2019). Different methods have been developed to gather scientific and technological knowledge and data for different purposes, but this information has only been general and focused solely on elementary processing operations, which do not take into account the entire processing operation. For example, Kansou et al. (2014) proposed a qualitative model of a unitary mixing operation using an expert system to predict the quality of wheat flour dough. Baudrit et al. (2015) modeled preharvest grape berry maturity—a critical characteristic for the wine industry—using expert knowledge and data and probabilistic graphical approaches. Belna et al. (2022) optimized microfiltration unit operation to integrate conflicting stakeholder objectives, such as maximizing product output quality while minimizing cost inputs and addressing environmental impacts. Baudrit et al. (2022) used data from scientific articles describing the entire milk microfiltration process including several unit operations in addition to the milk microfiltration step as skimming, heat treatment or storage. Those data are available in Buche et al. (2021). But the method presented in Baudrit et al. (2022) only proposes to learn a predictive model of the milk microfiltration unit operation in large-scale operational conditions including different membranes.

TABLE 1 Assessment of the efficiency aggregation method.

Decision tree	Number of action assessments	Number of disagreements
Excessive dripping	36	2
Excessive acidification	28	1
Excessive salting	24	2
Total	88	5

In a circular economy context, Belaud et al. (2022) proposed a decision-support system to rank alternative lignocellulosic waste transformation processes based on knowledge engineering tools to compile experimental data to assess potential environmental impacts. Munch et al. (2021) and Munch et al. (2022) combined ontology, probabilistic models and linked open data to generate, through a reverse engineering approach, agricultural wastes well-suited for processing biocomposites for food packaging. In both cases, ontological models facilitated analysis of the impact of the entire processing operation on the end-product quality or on indicators associated with the process, but the approaches required gathering of a substantial set of numerical experimental data to obtain good results. These approaches are unsuitable to achieve the objectives outlined in this paper because they are over-demanding in terms of obtaining sufficient numerical data to represent the entire range of collective knowledge at the level of a given food chain.

Fault tree analysis (FTA), which targets fault event risk assessment (Baig et al., 2013), may 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 instance, in Pahasup-anan et al. (2021), the authors 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 which 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 branches of the tree 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.

Our core ontology, which defines the decision tree structure, is comparable to the tree structure used in FTA as we also represent a decision tree linking explanatory situations to a given situation of interest. However, our objectives are quite different (see Section 2.1). Our original contribution compared to FTA consists of: (i) 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; (ii) representing levers and associated technological actions to solve a situation of interest; (iii) representing the action efficiency based on individual experience; (iv) representing contextual criteria associated with recommendations to help filter recommendations in a multicriteria way; and (v) 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 is based on collecting and representing the collective technical know-how available for a given domain (company, food chain, etc.).

Our ontological model may be compared with the COOK ontology (Ghrab et al., 2017), the Core Ontology of Organization Know-How and Knowing-That. In COOK, Know-How is defined as the capacity/disposition to perform an action. The COOK 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 viewpoint, we could consider that in the MCDSS the Situation concept is a specialization of CollectiveKnow-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. On the other hand, we propose a more elaborate representation of the COOK Proposition concept as we represent expert reasoning using the notion of causality between situations, the efficiency of an action and associated contextual criteria. In conclusion, 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 is not present in COOK. The part of our ontology which more or less corresponds to the Knowing-That concept part of COOK is more detailed because we have proposed a complete and operational 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), i.e., preliminary research that gave rise to the study described in this paper, several new contributions are proposed and assessed: (i) a model to aggregate action efficiency based on individual experience; (ii) an extended version of the ontology expressed in OWL, including the representation of action efficiency and associated information (key and complementary criteria); and (iii) the definition and implementation of the action view.

4.2. Key contributions

Here we showcased a new multicriteria decision-support system based on collective know-how in food chains to enhance food quality during the production process. This MCDSS is currently being used in production conditions in 13 AOP cheesemaking chain organizations involving professional stakeholders (cheese producers, experts working in technical centers) and professors from technological schools (ENILs), thereby comprising more than 60 users. Note that the model used to represent the decision trees may be refined in a flexible

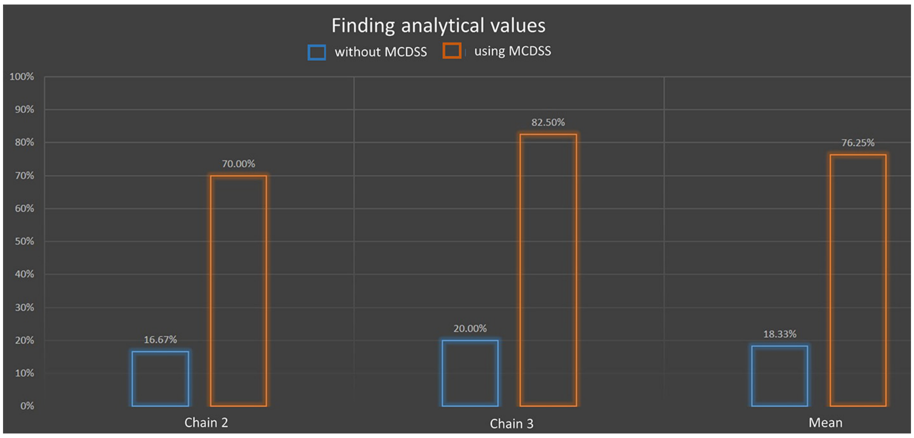


FIGURE 12
Test results (blue without MCDSS, brown with MCDSS) for analytical value findings: the y-axis represents the percentage of correct answers (mean value) in two chains and on average.

way until the level of detail sought by the expert is reached. This flexibility has been successfully used in the project as professors from technological schools have created generic cheese-making decision-trees. Operators introducing new chains in the project have tailored these generic decision trees to their specific cheese-making process.

Most of the MCDSS functionalities implemented to fulfill the specifications have been assessed, with promising results overall. All of the ontological model concepts are required to implement the MCDSS specifications. The choice to represent the causality relation between situations may be discussed. Indeed, a *HasCausality* semantic relation between two *Situation* nodes would have also been possible and simpler. However, as presented below in the perspectives of this work, it would be useful for advanced users conducting statistical analyses to be able to qualify the causality relation between situations in a future version of the MCDSS. By example, we would like to distinguish between types of sources, which contain the statistical analysis (internal study, bibliographical study). This metadatum should be associated with Causality nodes.

This justifies the modeling choice to set the stage for this future development. The comparison with and without the MCDSS was important to convince technicians of the MCDSS relevance. There are currently many obstacles to innovation, especially when it comes to know-how of which experts are sometimes afraid of losing (loss of employment, of power, etc.). Moreover, the direct use of mind maps without MCDSS assistance could be questioned. From our viewpoint the relevance is limited since the use of a simple mind map does not allow for three kinds of computerized analysis:

- The first corresponds to numerical aggregation of individual action efficiency assessments (see Specification 4). It provides an aggregated indicator, which is discussed and validated collectively by the technician team to determine the final action efficiency value, especially in the event of major disagreement. This value enriches the decision tree associated with a given situation of interest registered in the knowledge base.
- The second consists of the action multicriteria filtering mechanism generated by MCDSS queries, which reduces the list of candidate actions for a given situation of interest (see Specification 4).
- The third corresponds to the computing of complementary answers to enable assessment of the potential side effects with other actions and situations of interest (see Specification 3).

A typical use case of the MCDSS, which illustrates the relevance of those computerized analysis, is the following. A cheese maker has a problem of excessive salting. He/she queries the MCDSS on his/her phone using the *excessive salting* decision tree. First, using the Action view (see Figure 7), he/she selects the three corrective actions, which are very effective with a good/average agreement between experts of the chain. Clicking on the button identifying possible side effects (see Figure 8), he/she understands that using the *Brine duration* lever could be risky as four other defects may appear. Therefore, he/she navigates in the Explanatory view (see Figure 6) to compare the two remaining recommended actions (acidify or dilute the brine) will be able to verify if the analytical value associated with the situation *High brine density* corresponds to his/her actual situation to choose the action he/she will use.

Collected know-how consistency checking is consolidated and enriched throughout the workflow presented in Figure 1. First note that Buche et al. (2019) proposed a method that enables collective mind mapping dedicated to this MCDSS. Consequently, decision trees which are outcomes of this collective mind mapping activity are already validated as they contain the consolidated knowledge of the food chain experts resulting from collective discussion. Secondly, it is possible to verify that recommended actions are relevant by checking the technicians' feedbacks (Specification 4) after implementation of the recommendations. If the actions are good, then the recommendation remains valid. Conversely, further investigation may be required to understand why a recommended action failed. Thirdly, criteria associated with action efficiency assessments may contain information to verify action's relevance. For instance, an advanced chain has defined two criteria using the MCDSS: (1) StatisticalResults (yes/no), meaning that statistical results validating the recommendation have been obtained in the food chain; (2) BibliographicalResults (yes/no), meaning that results published in a scientific paper have validated the recommendation. In both cases, a link to a complementary website may be embedded in the decision tree branch to provide more information. Fourthly, it is possible to determine if certain suggestions delivered by the MCDSS were not executed. Indeed, this kind of action may be identified if the *Never* modality is associated with the "*number of action implementations*" information (*hasForObservations* property associated with the *Efficiency* concept). Explanations may be provided by analyzing values associated with the *ConsequenceCriteria* associated with the *Efficiency* concept. For instance, a given action has never been executed because it could generate a sanitary risk or be costly to implement. Fifthly, the collective mind mapping activity conducted to create and maintain decision trees may identify knowledge gaps. This means that in a given situation the experts may not know which action to recommend or may disagree on the action to recommend. Sometimes, they may know the action to recommend to solve a given situation of interest, without being able to explain why. In all of those cases, new experiments may be conducted to unlock knowledge gaps. Learning improvement in novel knowledge gap cases is a natural outcome of the collective mind mapping activity, which is the first step of the MCDSS workflow process.

In terms of upscaling, more advanced chains manage around a 100 decision trees and they will certainly increase to several hundreds. But big data are not involved and no scalability problems in terms of volumes should arise because we only represent expert knowledge. The problem would arise if we were to seek to represent the numerical experimental data so as to be able to create/assess this expert knowledge. But this is beyond the scope of this work.

5. Conclusion and perspectives

The perspectives of this project are numerous. In AOP cheesemaking chain organizations, the priority has been toward recommendations to correct organoleptic defects. But in the future the method will enable us to create decision trees to recommend actions to solve food safety problems or to achieve a given food quality. Moreover, we will extend know-how representation to the upstream part of the chain, including milk production. New methodological challenges will be tackled to take spatio-temporal knowledge representation into account. Advanced users who conduct statistical analyses may like to be able to qualify the causality relation between situations in a future version of the MCDSS. This extension will be easy to design thanks to the choice made in the ontology model to represent causality relations. Another prospect will be to take new MCDSS sustainability criteria into account. We will focus specifically on the environmental impact of cheese production. Using Semantic Web languages to implement the knowledge base will facilitate interoperability management with new sources of information that are also managed with those languages (Pénicaud et al., 2019; Cortesi et al., 2022a,b).

This MCDSS is a generic tool, which could potentially be used in different food and bio-product chains. Encouraging preliminary tests, as reported in Buche et al. (2019), have been conducted in the cereal (couscous) and dairy sectors (instant milk powder). Consequently, new dissemination activities will be conducted in the future.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://doi.org/10.57745/SEJP1B.

Ethics statement

Ethical approval was not required for the study involving human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants in accordance with the national legislation and the institutional requirements.

Author contributions

PB, SD, JCu, AO, and JCo contributed to conception and design of the study. JCu and AO organized the database. PB and JCo performed the statistical analysis. PB wrote the first draft of the manuscript. PB and SD wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This study has been partially financed by the French Ministry of Agriculture through the CASDAR system in the framework of the DOCaMEx project, the France Relance project Docamex an innovative digital tool at the service of cheese knowledge and know-how (12-363-DNUM-CAAF-003), and the French National Research Agency (ANR) in the framework of the D2KAB project (ANR-18-CE23-0017).

Acknowledgments

The authors thank experts from ENIL schools and technical centers devoted to AOP cheesemaking chains in France, namely Comté, Reblochon, Emmental de Savoie, Cantal, and Salers, for their help in the design and assessment of the methods and tools presented in this paper. We are especially grateful to Manon Recour (Syndicat Interprofessionnel du Reblochon) and Virginie Cucheval (Center Technique des Fromages Comtois) for the very rich exchanges throughout the MCDSS design process. We also thank Jérôme Fortin and David Manley who carefully read and comment this paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

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. *Innov. Food Sci. Emerg. Technol.* 46, 54–64. doi: 10.1016/j.ifset.2017.10.006

Baig, A., Ruzli, R., and Buang, A. (2013). Reliability analysis using fault tree analysis: A review. *Int. J. Chem. Eng. Appl.* 4, 169–173. doi: 10.7763/IJCEA.2013.V4.287

Baudrit, C., Buche, P., Leconte, N., Belna, M., Fernandez, C., and Gesan-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. *Int. J. Agric. Environ. Inf. Syst.* 13, 1–22. doi: 10.4018/IJAEIS.309136

Baudrit, C., Perrot, N., Brousset, J.-M., Abbal, P., Guillemin, H., Perret, B., et al. (2015). A probabilistic graphical model for describing the grape berry maturity. *Comput. Electr. Agri.* 118, 124–135. doi: 10.1016/j.compag.2015.08.019

Belaud, J. P., Prioux, N., Vialle, C., Buche, P., Destercke, S., Barakat, A., et al. (2022). Intensive data and knowledge-driven approach for sustainability analysis: Application to lignocellulosic waste valorization processes. *Waste Biomass Valorizat.* 13, 583–598. doi: 10.1007/s12649-021-01509-8

Belna, M., Ndiaye, A., Taillandier, F., Fernandez, C., Agabriel, L., and Gésan-Guiziou, G. (2022). Multiobjective optimization of skim milk microfiltration based on expert knowledge. *Expert Syst. Appl.* 205, 117624. doi: 10.1016/j.eswa.2022.117624

Buche, P., Cufi, J., and Oudot, A. (2022). *Food Quality Decision Tree Based on Collective Know-How (Capex Ontology)*. doi: 10.57745/SEJP1B

Buche, P., Cuq, B., Fortin, J., and Sipieter, C. (2019). Expertise-based decision support for managing food quality in agri-food companies. *Comput. Electr. Agri.* 163, 104843. doi: 10.1016/j.compag.2019.05.052

Buche, P., Dervaux, S., Leconte, N., Belna, M., Granger-Delacroix, M., Garnier-Lambrouin, F., et al. (2021). Milk microfiltration process dataset annotated from a collection of scientific papers. *Data Brief* 36, 107063. doi: 10.1016/j.dib.2021.107063

Chein, M., and Mugnier, M.-L. (2009). *Graph-Based Knowledge Representation and Reasoning*. London: Computational Foundations of Conceptual Graphs. Springer, Advanced Information and Knowledge Processing Series.

Cortesi, A., Dijoux, L., Yannou-Le Bris, G., and Pénicaud, C. (2022a). Data related to the life cycle assessment of 44 artisanally produced french protected designation of origin (PDO) cheeses. *Data Brief* 43, 108403. doi: 10.1016/j.dib.2022.108403

Cortesi, A., Dijoux, L., Yannou-Le Bris, G., and Pénicaud, C. (2022b). Explaining the differences between the environmental impacts of 44 French Artisanal Cheeses. *Sustainability* 14, 9484. doi: 10.3390/su14159484

Ghrab, S., Saad, I., Kassel, G., and Gargouri, F. (2017). A core ontology of know-how and knowing-that for improving knowledge sharing and decision making in the digital age. *J. Decision Syst.* 16, 138–151. doi: 10.1080/12460125.2016.1252231

Kansou, K., Chiron, H., Della Valle, G., Ndiaye, A., and Roussel, P. (2014). Predicting the quality of wheat flour dough at mixing using an expert system. *Food Res. Int.* 64, 772–782. doi: 10.1016/j.foodres.2014.08.007

Kim, D. H., Choc, W. I., and 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. doi: 10.1016/j.foodcont.2019.107019

Munch, M., Buche, P., Dervaux, S., Dibie, J., Ibanescu, L., Manfredotti, C., et al. (2022). Combining ontology and probabilistic models for the design of bio-based product transformation processes. *Expert Syst. Appl.* 203, 117406. doi: 10.1016/j.eswa.2022.117406

Munch, M., Buche, P., Manfredotti, C., Wuillemin, P. H., and 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 *15th International Conference on Metadata and Semantics Research, Nov 2021*. Madrid. doi: 10.1007/978-3-030-98876-0_1

Pahasup-anan, T., Kreetachat, T., Ruengphrathuengsuka, W., Wongcharee, S., Usahanunth, N., Imman, S., et al. (2021). Dust explosion risk assessment of extruded food production process by fault tree analysis. *ACS Chem. Health Saf.* 29, 91–97. doi: 10.1021/acs.chas.1c00036

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 PO^2 ontology. *Int. Dairy J.* 92, 3. doi: 10.1016/j.idairyj.2019.01.003

Sowa, J. F. (1984). *Conceptual Structures: Information Proc.700 in Mind and Machine*. Boston, MA: Addison–Wesley.

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 Eng. Rev.* 11, 44–60. doi: 10.1007/s12393-018-9186-x

Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 13 June 2023
DOI 10.3389/frai.2023.1201002

Using knowledge graphs to infer gene expression in plants

Anne E. Thessen[1]*, Laurel Cooper [2], Tyson L. Swetnam [3], Harshad Hegde [4], Justin Reese [4], Justin Elser [2] and Pankaj Jaiswal [2]

[1]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, United States, [2]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, United States, [3]BIO5 Institute, University of Arizona, Tucson, AZ, United States, [4]Environmental Genomics and Systems Biology Division, Berkeley Lab (DOE), Berkeley, CA, United States

OPEN ACCESS

EDITED BY
Marie-Angélique Laporte,
Bioversity International, France

REVIEWED BY
Ashni Sedani,
Accenture, United Kingdom
Angela Kranz,
Helmholtz Association of German Research Centres (HZ), Germany

*CORRESPONDENCE
Anne E. Thessen
✉ annethessen@gmail.com

RECEIVED 05 April 2023
ACCEPTED 23 May 2023
PUBLISHED 13 June 2023

CITATION
Thessen AE, Cooper L, Swetnam TL, Hegde H, Reese J, Elser J and Jaiswal P (2023) Using knowledge graphs to infer gene expression in plants. *Front. Artif. Intell.* 6:1201002.
doi: 10.3389/frai.2023.1201002

Introduction: Climate change is already affecting ecosystems around the world and forcing us to adapt to meet societal needs. The speed with which climate change is progressing necessitates a massive scaling up of the number of species with understood genotype-environment-phenotype (G×E×P) dynamics in order to increase ecosystem and agriculture resilience. An important part of predicting phenotype is understanding the complex gene regulatory networks present in organisms. Previous work has demonstrated that knowledge about one species can be applied to another using ontologically-supported knowledge bases that exploit homologous structures and homologous genes. These types of structures that can apply knowledge about one species to another have the potential to enable the massive scaling up that is needed through *in silico* experimentation.

Methods: We developed one such structure, a knowledge graph (KG) using information from Planteome and the EMBL-EBI Expression Atlas that connects gene expression, molecular interactions, functions, and pathways to homology-based gene annotations. Our preliminary analysis uses data from gene expression studies in *Arabidopsis thaliana* and *Populus trichocarpa* plants exposed to drought conditions.

Results: A graph query identified 16 pairs of homologous genes in these two taxa, some of which show opposite patterns of gene expression in response to drought. As expected, analysis of the upstream cis-regulatory region of these genes revealed that homologs with similar expression behavior had conserved cis-regulatory regions and potential interaction with similar trans-elements, unlike homologs that changed their expression in opposite ways.

Discussion: This suggests that even though the homologous pairs share common ancestry and functional roles, predicting expression and phenotype through homology inference needs careful consideration of integrating cis and trans-regulatory components in the curated and inferred knowledge graph.

KEYWORDS

knowledge graph (KG), plant genome, gene expression, ontology, phenotype

Introduction

Climate change is already affecting ecosystems around the world and forcing us to explore ways to adapt to meet societal needs. This is particularly true in crop science where researchers are working to identify and predict genes and their resulting phenotypes under different environmental conditions in order to secure food production under a new climate regime (Thudi et al., 2021; Tian et al., 2021). Understanding gene/phenotype/environment relationships requires a large data set which can be difficult to collect, so most researchers focus on a small number of heavily studied species. The speed with which climate change is progressing necessitates a massive scaling up of the number of species with understood G/P/E dynamics. The research and the knowledge gained in this area will also help human exploration in space, where plants will play an important role (Barker et al., 2023). Previous study has demonstrated that knowledge about one species can be applied to another using ontologically supported knowledgebases that exploit homologous structures and orthologous genes (Naithani et al., 2020). These types of knowledge structures that can apply knowledge about one species to another have the potential to enable the massive scaling up that is needed.

An important part of predicting phenotype is understanding the complex gene regulatory networks present in plants. This study will focus on the promoter region, the 5′ cis-regulatory regions of the homologs. This region is a portion of the DNA strand that is "upstream" from the 5′ end of the gene's coding start site and provides selective binding sites for trans-acting factors such as transcription factors, repressors, and activators that regulate the expression of the gene (Liu et al., 1999). These regions are just one element of the gene expression process. Studying the expression of trans-acting factors is important for understanding the spatiotemporal dynamics of molecular interactions that help adapt or overcome stress. Resources such as the Gene Ontology (The GO Consortium, 2021), Planteome (Cooper et al., 2018), Plant Reactome (Naithani et al., 2020), and KnetMiner (Hassani-Pak et al., 2021) contain much of what we know about gene function, gene regulatory networks, and phenotypes in the form of Gene X regulates Gene Y and Gene Y impacts phenotype Z, but the contextual effect of environmental conditions under which these interactions happen is almost always not included in the annotations. Not all plants and their genes are characterized in detail, but if it is included, the environmental context is usually detailed only in the metadata. Investigations that use protein domain identification and gene homology-based methods to infer the functional role a gene carries out in a given species may be overlooking the spatial and temporal dynamics of mRNA expression that determines whether a gene product (protein) will be present at the desired time and place to serve a molecular function. The interactive nature of genes, environments, and phenotypes requires a data structure that can represent qualitative relationships (e.g., "has phenotype" or "regulates") and integrate heterogeneous data types in a single, queryable framework. One of these data structures is a knowledge graph (KG) (Sheth et al., 2019).

A graph is made up of objects (nodes) and the relationships (edges) between those objects and, in this context, represents what we know about how biological and environmental entities (objects) interact. Rather than store data in a table or database, a knowledge graph stores the synthesized knowledge we gain from the data, e.g., Gene X has phenotype Y. As more knowledge is added to the graph, more complex queries, network analyses, and inferences can be made. Important examples include the use of knowledge graphs in rare disease diagnosis in humans (Zemojtel et al., 2014), drug repurposing (Reese et al., 2021), improving cancer treatment (Gogleva et al., 2022), and meta-analyses (Tiddi et al., 2020). KGs used for translational science rarely contain environmental exposures even though we know environmental conditions are an important part of gene expression dynamics. The exact way to model exposures in a KG is still under development (Chan et al., 2023). A KG containing information about plant genomics and phenomics under different environmental conditions can be used to generate hypotheses *in silico* for targeting, thereby reducing the number of *in vivo* experiments that need to be conducted, saving time and resources.

This study examines gene expression patterns in response to drought conditions in four plant species, such as *Arabidopsis thaliana*, *Zea mays*, *Sorghum bicolor*, and *Populus trichocarpa*. The central motivation of this study is to assess the feasibility of using homologs to make predictions about gene expression in multiple species.

Materials and methods

Data description

Planteome

The Planteome (https://planteome.org/) is a centralized web portal with a suite of interrelated ontologies for plants and a database of plant genomics data, annotated to the ontology terms (Cooper et al., 2018). In the October 2020 release (version 4.0), the Planteome database included approximately 60,000 ontology terms and more than 3 million data objects, which are connected to ontology terms through approximately 20 million associations. The Planteome database has plant genomic information covering 125 plant taxa. The data available in the Planteome and annotated with ontology terms, include plant gene expression data, traits, phenotypes, genomes, and germplasm sources.

The ontologies developed in-house by the Planteome project include the Plant Ontology (PO; Cooper et al., 2018; Walls et al., 2019), which describes plant anatomical structures and developmental stages, the Plant Trait Ontology (TO) for traits and phenotypes, and the Plant Experimental Conditions Ontology (PECO), which describes experimental conditions and plant exposures. In addition to these, the Planteome hosts the collaborator reference ontologies—the Gene Ontology (GO; The GO Consortium, 2021), Phenotype and Trait Ontology (PATO; Gkoutos et al., 2018), and also a number of species-specific trait dictionaries developed by the Crop Ontology (CO; Shrestha et al., 2010; Arnaud et al., 2020). In the current release, the Planetome includes 11 of the CO trait dictionaries, mapped to the TO.

GO annotations were computationally generated for new species using InParanoid and InterProScan (Shulaev et al., 2011; Myburg et al., 2014). InParanoid was used to predict gene orthology based on the *Arabidopsis thaliana* associations generated by TAIR (Reiser et al., 2022). InterProScan was used to add GO annotations to genes via inference by analyzing protein families and domain mappings (Paysan-Lafosse et al., 2023).

EMBL-EBI expression atlas

The EMBL-EBI Expression Atlas (GXA) can be accessed online and is part of the European Bioinformatics Institute (Papatheodorou et al., 2020). It contains manually curated and analyzed data from over 900 plant experiments that have been re-analyzed using the latest versions of the reference plant genome assembly and annotations and by deploying a standardized analysis workflow. Every experiment is fully documented with metadata and provenance.

Gene expression data were downloaded as a table from GXA after searching for desired species and environmental conditions. Data were filtered to include only genes that had statistically different gene expressions ($p<0.05$) compared with a baseline that was <-1 or >1. Genes with positive differential expression were annotated as having increased expression. Genes with negative differential expression were annotated as having decreased expression. The tabulated data were annotated with additional ontology terms where appropriate and made available in GitHub for graph construction.

Creating the graph

The graph was created by combining data from Planteome, the GXA, PO, TO, GO, and PECO using the tools available at KG-Hub (Caufield et al., 2023). First, the data and mapping files were downloaded from their respective data repositories. GO-Basic and NCBI Tax-Slim were downloaded from the OBO Foundry in javascript object notation (JSON) format. PO and TO were downloaded from the OBO Foundry in owl format and transformed to JSON using ROBOT (Jackson et al., 2019). Data files containing information about *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, *Populus trichocarpa*, and *Arabidopsis thaliana* were downloaded from Planteome servers in GAF format. Data files containing differential gene expression data involving *Sorghum bicolor*, *Zea mays*, *Oryza sativa*, *Populus trichocarpa*, and *Arabidopsis thaliana* in drought and saline environments were downloaded from the GXA. Several mapping files were used to normalize gene and trait identifiers. Rice gene identifiers were mapped to *Oryza sativa* v7.0 using the ID converter file from the Rice Annotation Project Database (Ouyang et al., 2007; Sakai et al., 2013). Maize gene identifiers were mapped to Zm-B73-REFERENCE-NAM-5.0 assembly using a mapping file that includes all B73 assembly versions and includes the DAGchainer analysis which was obtained from MaizeGDB (Portwood et al., 2019; EMBL-EBI). Poplar gene identifiers were mapped to the reference genome using a mapping file from Gramene (Tello-Ruiz et al., 2018). *Sorghum* gene names were normalized to *Sorghum bicolor* v3.1.1 (McCormick et al., 2018). Plant traits and phenotypes were annotated with TO terms using a look-up dictionary file. Second, each of the data files was transformed into standardized nodes and edges in a tsv file using custom scripts. These scripts normalized gene and trait identifiers using ontologies and the provided mapping files and annotated every entity with a Biolink semantic type (Table 1), and relationships between entities were described using Biolink predicates (Table 2). The graph was assembled according to the Biolink model, which provides standard semantic types and relationships for biological entities (Unni et al., 2022).

There was not enough overlapping expression data to include *O. sativa* or saline environments in this analysis, but they were included in the graph.

TABLE 1 Identifiers and Biolink semantic types assigned to elements of the graph.

Biological element	Identifier	Biolink type
Plant part	PO	Anatomical entity
Growth stage	PO	Life stage
Plant trait	TO	Phenotypic feature
Zea mays gene	Zm00001eb IDs	Genomic entity
Sorghum bicolor gene	Sobic IDs	Genomic entity
Oryza sativa gene	LOC_Os IDs	Genomic entity
Populus trichocarpa gene	POPTR IDs	Genomic entity
Experimental condition	PECO	Environmental exposure
QTL	Gramene IDs	Genomic entity
Cultivar	NCBITaxonomy	Organismal entity
Taxon	NCBITaxonomy	Organism taxon
Cellular component	GO	Cellular component
Molecular function	GO	Molecular function
Biological process	GO	Biological process
Germplasm	GRIN and IRIC IDs	Organismal entity

TABLE 2 Edges and their Biolink predicates.

Subject entity	Predicate type	Object entity
Genomic entity	In taxon	Organism taxon
Genomic entity	Active in	Cellular component
Genomic entity	Regulates	Biological process
Genomic entity	Enables	Molecular function
Genomic entity	Expressed in	Anatomical entity
Genomic entity	Expressed in	Life stage
Genomic entity	Has phenotype	Phenotypic feature
Genomic entity	Orthologous to	Genomic entity
Organism taxon	Has phenotype	Phenotypic feature
Organismal entity	In taxon	Organism taxon
Organismal entity	Has phenotype	Phenotypic feature
Environmental exposure	Increases expression of	Genomic entity
Environmental exposure	Decreases expression of	Genomic entity

The third and final step merged the transformed tsv files into a deduplicated list of nodes and edges in KGX format. The final graph consisted of over 400,000 nodes and over 5,000,000 edges and contained additional data from EOLTraitbank that was not used in this study (Figure 1). Specific information about quantitative and qualitative plant phenotypes was represented as an edge property (Figure 2).

Querying the graph

The merged node file and edge file were uploaded into Neo4j for exploration and query. A Cipher query (Box 1) was used

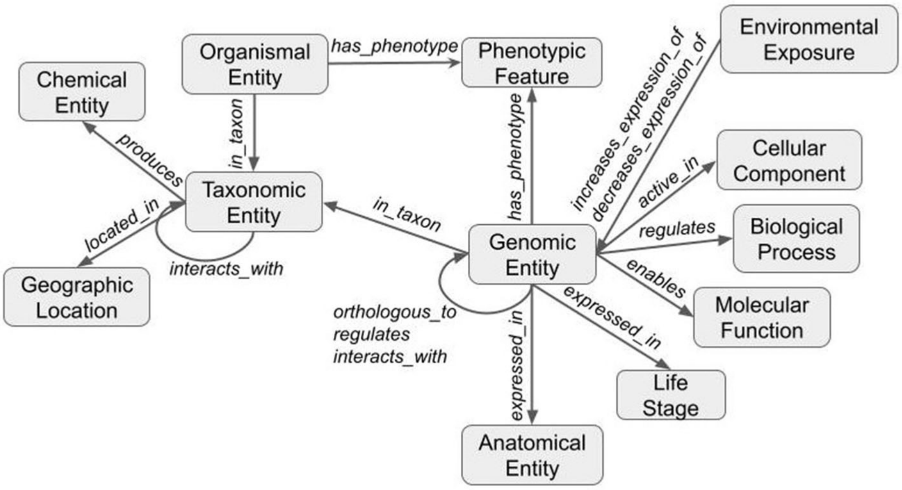


FIGURE 1
Structure of the knowledge graph. Data are transformed using ontologies and the Biolink model to form a graph. Nodes (gray boxes) are labeled with Biolink semantic type and edges (gray arrows) are labeled with Biolink predicate. Arrows indicate directionality.

TABLE 3 Gene expression in *A. thaliana* and *P. trichocarpa* homologous genes under drought conditions.

A. thaliana Gene*	*P. trichocarpa* Gene*	Gene function (from Planteome)
AT3G49960 ↓	POPTR_007G053400v3 ↑	Peroxidase activity, response to oxidative stress, heme binding
AT1G70710 ↓	POPTR_010G109200v3 ↓	Catalytic activity, hyrolase activity, carbohydrate metabolic process
AT1G10550 ↓	POPTR_014G115000v3 ↓	Xyloglucan metabolism, hyrolase activity, carbohydrate metabolic process, cell wall biogenesis
AT5G67400	POPTR_007G053400v3 ↑	Peroxidase activity, response to oxidative stress, heme binding, hydrogen peroxide catabolic process
AT5G23210 ↓	POPTR_005G091700v3 ↓	Proteolysis, serine-type carboxypeptidase activity
AT1G67750 ↓	POPTR_008G182200v3 ↓	Pectate lyase activity, metal ion binding
AT2G39530 ↓	POPTR_010G205300v3 ↑	iron/sulfur cluster binding
AT5G13140 ↓	POPTR_003G167100v3 ↓	Response to nematode, pectate lyase activity, metal ion binding
AT1G11580 ↓	POPTR_011G025400v3 ↑	Enzyme inhibitor activity, pectinesterase activity, cell wall modification, rRNA N-glycosylase activity, aspartyl esterase activity, toxin activity, defense response
AT3G27400 ↓	POPTR_001G339500v3 ↑	Response to nematode, pectate lyase activity, metal ion binding
AT4G02330 ↓	POPTR_014G127000v3 ↑	Enzyme inhibitor activity, pectinesterase activity, cell wall modification, response to stress, aspartyl esterase activity
AT5G20630 ↓	POPTR_006G142600v3 ↓	Manganese ion binding, nutrient reservoir activity
AT4G26260 ↑	POPTR_018G069700v3 ↓	Iron ion binding, inositol oxygenase activity, syncytium formation, L-ascorbic acid biosynthetic pathway
AT2G44990 ↑	POPTR_014G056800v3 ↓	Oxidoreductase activity, secondary shoot formation, carotene catabolic process, strigolactone biosynthetic process, xanthophyll catabolic process, metal ion binding
AT1G70710 ↓	POPTR_010G109200v3 ↓	Cellulase activity, cell wall modification, hydrolase activity
AT1G12940 ↑	POPTR_015G081500v3 ↓	Transmembrane transport

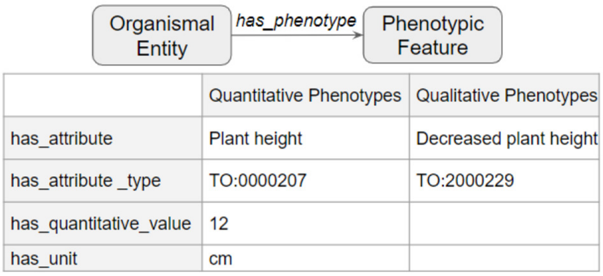
*↓ Indicates decreased expression and ↑ indicates increased expression.

to find all of the homologous genes that had been documented to have differential gene expression in either a drought or a saline environment (Supplementary material 1, 2). The saline environment did not return overlapping data.

Genes returned from the query for the drought environment were compared based on GO annotations (Supplementary material 3), but this also did not give enough data to make conclusions using PANTHER (Supplementary material 4).

Comparing promoter regions

We collected 5′-regulatory regions of the identified genes (700–900 bp) using BioMart in the Gramene database (Spooner et al., 2012) and searched for potential transcription factor-binding sites using PlantPAN (Chow et al., 2016). Using these data (Supplementary material 5), we created a matrix comparing the occurrence of each transcription factor in the binding site of each gene pair and made note of which were or were not held in common. We used ClustVis (Metsalu and Vilo, 2015) to examine the similarity between the transcription factor-binding sites for each of the *Populus* and *Arabidopsis* gene pairs using PCA. A total of 12 transcription factor-binding sites (AT-Hook, bHLH, C2H2, Dehydrin, Dof, GATA, Homeodomain, Myb/SANT, NF-YB, TBP, Trihelix, and ZF-HD) were present in the promoter regions of all the genes studied and thus were removed from clustering analysis. The same data were fed into Morpheus (Müller et al., 2008) for hierarchical clustering performed with default parameters using One minus Pearson's correlation and complete linkage methods on the TF-binding site annotations. Additional similarity matrices were created using Pearson's correlation metric to separately examine the TF-binding site annotations for genes with similar and contrasting expression profiles. The correlation heatmap colors were adjusted for visualization purposes.



	Quantitative Phenotypes	Qualitative Phenotypes
has_attribute	Plant height	Decreased plant height
has_attribute _type	TO:0000207	TO:2000229
has_quantitative_value	12	
has_unit	cm	

FIGURE 2
Phenotype data in edge properties. Detailed phenotype information was represented as a collection of edge properties that can accommodate quantitative and qualitative phenotypes.

Data availability

The merged KG data are hosted on the CyVerse DataCommons (https://datacommons.cyverse.org/browse/iplant/home/shared/genophenoenvo). The KG data are available for direct download

BOX 1 Cipher query.

MATCH (e {id:'PECO:0007404'})-[r]->(g),(g)-[q:'biolink: orthologous_to']-(h), (e {id:'PECO:0007404'})-[s]->(h) RETURN *

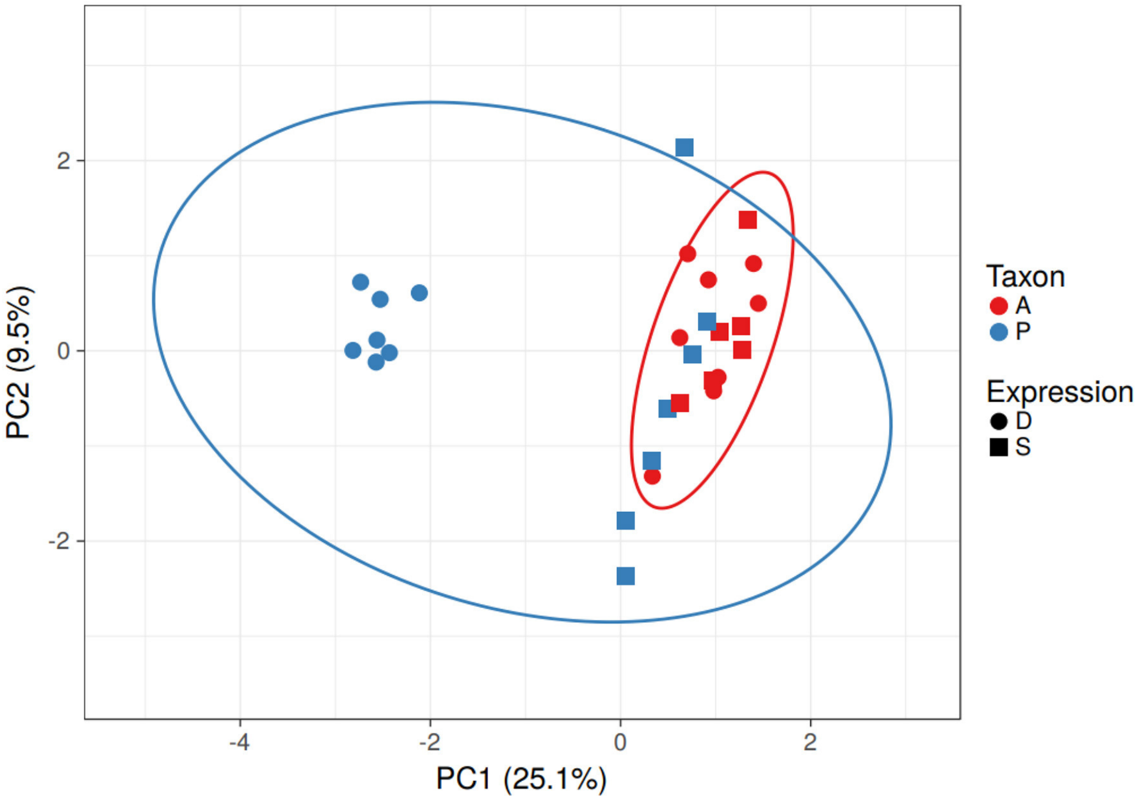


FIGURE 3
Clustering of *Populus* and *Arabidopsis* genes based on similarity of the transcription factor-binding sites in the promoter region - PCA. The *Populus* genes from the differentially expressed homolog pairs (blue circles) clustered away from the other *Populus* (blue) and *Arabidopsis* (red) genes. Differentially expressed genes are represented as circles and similarly expressed genes are represented as squares. Note that taxonomic differences (blue and red ovals) do not explain the differences in gene expression. No scaling is applied to rows; SVD with imputation is used to calculate principal components. X and Y axes show principal component 1 and principal component 2 that explain 25.1 and 9.5% of the total variance, respectively. *N* = 29 data points.

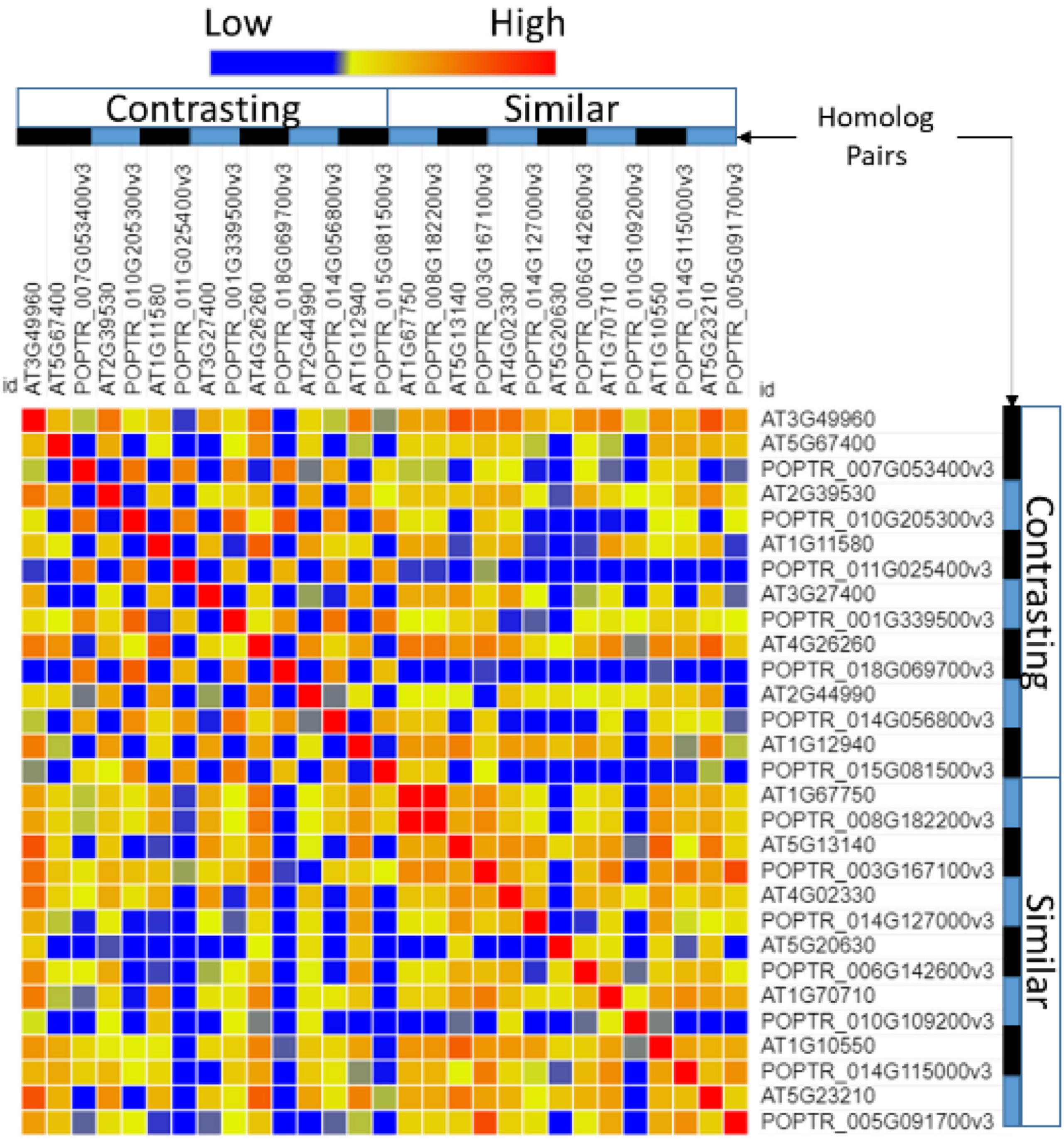


FIGURE 4
Similarity of the transcription factor-binding sites in the promoter region of *Populus* and *Arabidopsis* homologous gene pairs. Poplar (POPTR) and *Arabidopsis* (AT) genes were grouped into their homolog pairs and whether they had similar or contrasting gene expression when exposed to drought. This figure shows that the promoter regions of pairs with contrasting expressions were less similar (blue) and the promoter regions of pairs with similar expressions were more similar (red).

or remote visualization via CyVerse WebDav service (https://data.cyverse.org/dav-anon/iplant/commons/community_released/genophenoenvo/kg/) using visualization software such as Neo4J. The Python code used to create the graphs is publicly hosted on GitHub (https://github.com/genophenoenvo/knowledge-graph). The final merged KG includes two tab-separated value (tsv) files which include the edges and nodes.

Results

The graph query returned 62 pairs of homologous genes from *Sorghum bicolor*, *Zea mays*, *Arabidopsis thaliana*, and *Populus trichocarpa* (Supplementary material 6), but only 16 pairs between *A. thaliana* and *P. trichocarpa* had documented similar (8) and differential (8) expressions in drought conditions (Table 3). All of

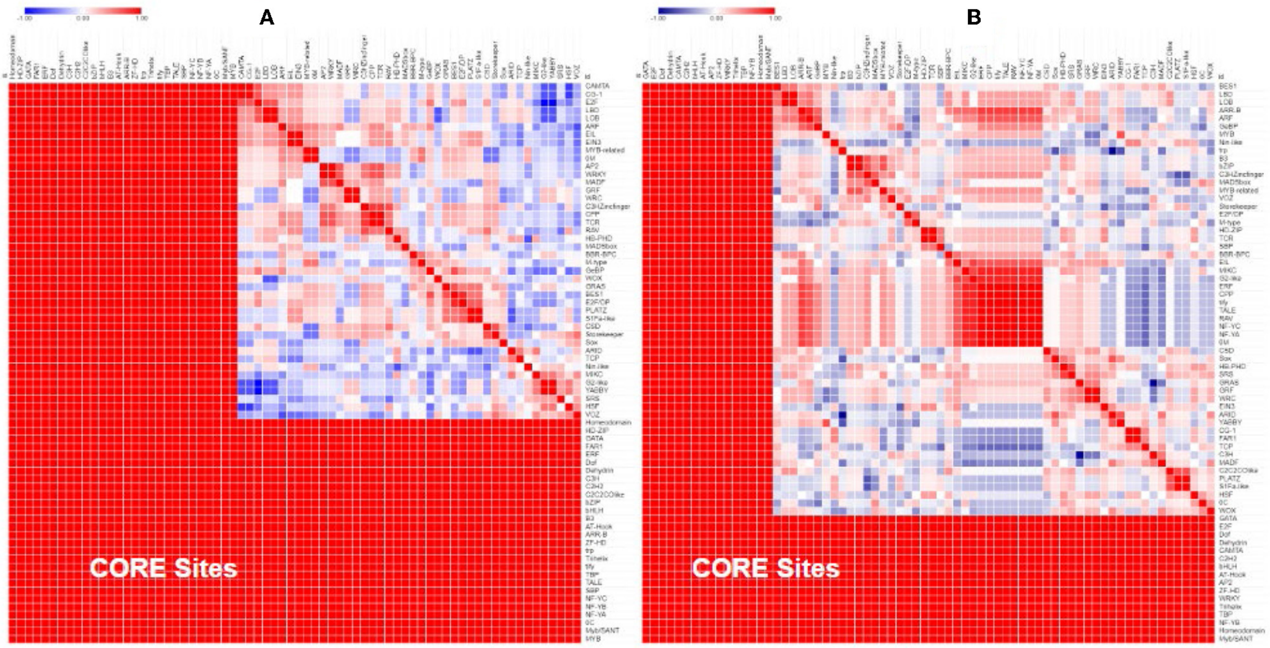


FIGURE 5
Similarity of the transcription factor-binding sites in the promoter region of *Populus* and *Arabidopsis* genes grouped by their expression profile. Genes that were similarly expressed in a drought treatment **(A)** had more similar promoter regions (red) than genes that were differentially expressed **(B)**.

the genes with similarly expressed pairs had decreased expression. Expression data for the 16 homologous pairs of *A. thaliana* and *P. trichocarpa* came from two studies in GXA (de Simone et al., 2017; Filichkin et al., 2018).

Based on the predicted transcription factor-binding sites in the promoter regions, the *Populus* genes in the differentially expressed homolog pairs cluster separately from the other *Populus* and *Arabidopsis* genes (Figure 3). This difference is driven by a group of 11 transcription factor-binding sites that are absent in the promoter regions of the subset of divergent *Populus* genes (RAV, MIKC, NAM, G2-like, CPP, ARR-B, tify, TALE, NF-YC, ERF, and NF-YA). The separation of these genes cannot be explained by the taxon or the study providing the data (which overlaps the taxon).

There were seven *Populus* genes that clustered away from the others. All but one (POPTR_014G056800v3 involved in strigolactone biosynthesis) were hypothetical proteins (According to Gramene). GO annotations for these genes clustered around transporter activity, catabolic activity, response to stress, binding, and catalytic activity. The 11 transcription factors absent in the binding sites of the *Populus* genes include proteins involved in plant stress response in *Arabidopsis* (According to UniProt).

A comparison of the promoter regions between homolog pairs showed that homologs that were expressed similarly had more similar promoter regions than pairs that were expressed differentially (Figure 4).

Separate comparisons of the promoter regions from gene pairs with contrasting expression profiles also show that gene pairs with similar expression had more similar promoter regions (Figure 5A) and gene pairs with contrasting expression had less similar promoter regions (Figure 5B).

Discussion

This study shows that one can use *in silico* experiments to predict gene expression in drought conditions using homologous gene families in some species pairs but not all. This study supports previous findings that in some cases, promoter regions evolve separately from the coding region of the genes they regulate (Tirosh et al., 2008). Thus, we can translate knowledge about gene expression in one species to another, but we need to include these dynamics in the data infrastructures we use to make this translation, in this case, KGs. Many data structures link a gene to a phenotype, trait, or disease without specific expression information. The current representation of differential gene expression links exposure to a chemical or a drug to the increased or decreased expression of a specific gene in the context of toxicology and drug development (Fecho et al., 2022; Unni et al., 2022). Gene regulatory networks are represented as mini-networks of genes that influence other genes (The GO Consortium, 2021), but many of these networks are still unknown in plants. In the short term, *in silico* KG experiments involving gene expression can be improved by including empirically validated gene expression patterns of homologs.

Gene regulatory networks in plants have been developed using a combination of experimental and computational approaches (Kulkarni and Vandepoele, 2020). Methods combining high-throughput DNA sequencing (ChIP-seq) and expression data have successfully revealed the detailed regulatory networks controlling flowering (Chen et al., 2018) but are difficult to scale. Methods such as ATAC-seq and DAP-seq are more scalable but only reveal a partial picture of the regulatory network (O'Malley et al., 2016;

Maher et al., 2018). KGs can be used to infer regulatory networks at scale, but the quality is highly dependent on the data used to build the KG. The advantage of applying a KG is the ability to integrate incredibly heterogeneous data in a single graph, thus modeling regulatory networks in their larger biological context. An example of this application is the relatively new field of "network medicine" that uses KGs to examine the progression of disease (Silverman et al., 2020). The main disadvantage of KGs in this application is that large amounts of computable data and domain-specific knowledge models are needed to create a graph of this type. Many disciplines do not have these resources available. While KGs can infer gene regulatory networks, these networks should always be confirmed using established experimental and computational approaches.

These opposing gene expression patterns are not a concern for researchers who are only interested in finding a list of genes that are potentially important in a specific context. It is not until one needs to generate hypotheses about the impact of the environment on the biological function that more complex graph representations become needed. If we are to incorporate the effect of the environment, we need to know more than that Gene X has phenotype Y. We need to know if the environmental effect increases or decreases the expression of the gene and the biological consequences of that change in expression. In some cases, we may only know that an environment is linked to a specific phenotype without knowing the underlying mechanism. This information can still add useful knowledge to the graph. In some cases, the graph itself can be used to generate hypotheses about the interplay between genes, biological processes, molecular functions, cellular components, and an observed phenotype.

Despite having the graph available to quickly explore the data and locate genes of interest, the workflow for comparing the promoter regions required substantial manual intervention. In this instance, we only had 16 gene pairs to explore, but scaling up these types of analyses will require the ability to traverse data annotated with gene identifiers and gene coordinates. Future studies should include extending the graph model to include these data types.

The semantic representation of the effect of environmental exposure on gene expression is more straightforward for the effects of a chemical or a substance, such as phenol or rubber cement. Data can be collected in the laboratory using model organisms, and the results added to the graph for analysis and translational research. Everyday environmental exposures are rarely this simple and frequently involve exposure to many types of substances in different contexts, such as climate or socioeconomic status. Future studies may need to develop ontologies and semantic representations for these more complex exposures.

Our observations support our hypothesis and justify the extension of our KG to include TF-binding site annotations and the actual TF genes, which are either known empirically or are supported by co-expression network analysis. In future, an investigation of conservation vs. non-conservation of cis- and trans-regulatory regions of genes may improve the understanding of interspecies and intraspecies responses to stress and adaptation.

Data availability statement

The merged KG data are hosted on the CyVerse DataCommons (https://datacommons.cyverse.org/browse/iplant/home/shared/genophenoenvo). The KG data are available for direct download or for remote visualization via CyVerse WebDav service (https://data.cyverse.org/dav-anon/iplant/commons/community_released/genophenoenvo/kg/) using visualization software such as Neo4J. The python code used to create the graphs are publicly hosted on GitHub (https://github.com/genophenoenvo/knowledge-graph). The final merged KG includes two tab-separated value (tsv) files which include the edges and nodes.

Author contributions

AT developed and framed research question(s), analyzed data, contributed to data analysis, developed software, contributed to writing and revising the paper, and project administration and management. HH contributed to data analysis. TS contributed to data analysis and contributed to writing and revising the paper. JR developed software, validated results or software, developed and framed research question(s), and contributed to writing and revising the paper. LC contributed to data analysis, project administration and management, and contributed to writing and revising the paper. PJ developed and framed research question(s), analyzed data, contributed to data analysis, and contributed to writing and revising the paper. JE contributed to data analysis, developed software, validated results or software, and contributed to writing and revising the paper. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the National Science Foundation grant awards #1940330, #1939945, #1940059, and #1940062. CyVerse is based upon work supported by the National Science Foundation grant awards #0735191, #1265383, and #1743442.

Acknowledgments

The authors would like to thank Sierra Moxon and Harry Caufield for their technical assistance and helpful conversation. The preprint for this paper can be found at https://doi.org/10.5281/zenodo.7764762.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arnaud, E., Hazekamp, T., Laporte, M. A., and Antezana, E. (2020). *Crop Ontology Governance and Stewardship Framework*. Available online at: https://cgspace.cgiar.org/handle/10568/118001 (accessed May 31, 2023).

Barker, R., Kruse, C. P. S., Johnson, C., Saravia-Butler, A., Fogle, H., and Chang, H-S., et al. (2023). Meta-analysis of the space flight and microgravity response of the arabidopsis plant transcriptome. *NPJ Microgr.* 9, 21. doi: 10.1038/s41526-023-00247-6

Caufield, J. H., Putman, T., Schaper, K., Unni, D. R., Hegde, H., Callahan, T. J., et al. (2023). KG-Hub – building and exchanging biological knowledge graphs. arXiv. Available online at: http://arxiv.org/abs/2302.10800

Chan, L. E., Thessen, A. E., Duncan, W. D., Matentzoglu, N., Schmitt, C., Grondin, C. J., et al. (2023). The Environmental Conditions, Treatments, and Exposures Ontology (ECTO): Connecting Toxicology and Exposure to Human Health and beyond. *J. Biomed. Semantics.* 14, 3. doi: 10.1186/s13326-023-00283-x

Chen, D., Yan, W., Fu, L. Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. *Nat. Commun.* 9, 4534. doi: 10.1038/s41467-018-06772-3

Chow, C. N., Zheng, H. Q., Wu, N. Y., Chien, C. H., Huang, H. D., Lee, T. Y., et al. (2016). PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. *Nucleic Acids Res.* 44, D1154–D1160. doi: 10.1093/nar/gkv1035

Cooper, L., Meier, A., Laporte, M-A., Elser, J. L., and Mungall, C., Sinn, B. T., et al. (2018). The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.* 46, D1168–D1180. doi: 10.1093/nar/gkx1152

de Simone, A., Hubbard, R., Torre, N. V., Velappan, Y., Wilson, M., Considine, M. J., et al. (2017). Redox changes during the cell cycle in the embryonic root meristem of *Arabidopsis Thaliana*. *Antioxid Redox Signal.* 27, 1505–1519. doi: 10.1089/ars.2016.6959

EMBL-EBI. n.d. *ENA Browser*. Available online at: https://www.ebi.ac.uk/ena/browser/view/GCA_902167145.1 (accessed March 10, 2023).

Fecho, K., Thessen, A. E., Baranzini, S. E., Bizon, C., Hadlock, J. J., Huang, S., et al. (2022). Progress toward a Universal Biomedical Data Translator. *Clin. Transl. Sci.* 15, 1838–1847. doi: 10.1111/cts.13301

Filichkin, S. A., Hamilton, M., Dharmawardhana, P. D., Singh, S. K., Sullivan, C., Ben-Hur, A., et al. (2018). Abiotic stresses modulate landscape of poplar transcriptome via alternative splicing, differential intron retention, and isoform ratio switching. *Front. Plant Sci.* 9, 5. doi: 10.3389/fpls.2018.00005

Gkoutos, G. V., Schofield, P. N., and Hoehndorf, R. (2018). The anatomy of phenotype ontologies: Principles, properties and applications. *Brief. Bioinform.* 19, 1008–1021. doi: 10.1093/bib/bbx035

Gogleva, A., Polychronopoulos, D., Pfeifer, M., Poroshin, V., Ughetto, M., Martin, M. J., et al. (2022). Knowledge graph-based recommendation framework identifies drivers of resistance in EGFR mutant non-small cell lung cancer. *Nat. Commun.* 13, 1667. doi: 10.1038/s41467-022-29292-7

Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., et al. (2021). KnetMiner: A comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol. J.* 19, 1670–1678. doi: 10.1111/pbi.13583

Jackson, R. C., Balhoff, J. P., Douglass, E., Harris, N. L., Mungall, C. J., Overton, J. A. R. O. B. O. T., et al. (2019). A tool for automating ontology workflows. *BMC Bioinformatics.* 20, 407. doi: 10.1186/s12859-019-3002-3

Kulkarni, S. R., and Vandepoele, K. (2020). Inference of plant gene regulatory networks using data driven methods: A practical overview. *Gene Regul. Mecha.* 1863, 194447. doi: 10.1016/j.bbagrm.2019.194447

Liu, L., White, M. J., and MacRae, T. H. (1999). Transcription factors and their genes in higher plants functional domains, evolution and regulation. *Eur. J. Biochem.* 262, 247–257. doi: 10.1046/j.1432-1327.1999.00349.x

Maher, K. A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D. A., et al. (2018). Profiling of accessible chromatic regions across multiple plant species and cell types reveals common gene regulatory principles and new control modules. *Plant Cell.* 30, 15–36. doi: 10.1105/tpc.17.00581

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. *Nucleic Acids Res.* 43, W566–W570. doi: 10.1093/nar/gkv468

Müller, E., Assent, I., Krieger, R., Jansen, T., and Seidl, T. (2008). "Morpheus," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA: ACM). doi: 10.1145/1401890.1402026

Myburg, A. A., Grattapaglia, D., Tuskan, G. A., Hellsten, U., Hayes, R. D., Grimwood, J., et al. (2014). The genome of eucalyptus grandis. *Nature.* 510, 356–362. doi: 10.1038/nature13308

Naithani, S., Gupta, P., Preece, J., D'Eustachio, P., Elser, J. L., Garg, P., et al. (2020). Plant reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* 48, D1093–1103. doi: 10.1093/nar/gkz996

O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., et al. (2016). Cistrome and epicistrome features shape the regulatory DNA landscape. *Cell.* 166, 1598. doi: 10.1016/j.cell.2016.08.063

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.* 35, D883–D887. doi: 10.1093/nar/gkl976

Papatheodorou, I., Moreno, P., Manning, J., Muñoz-Pomer Fuentes, A., George, N., Fexova, S., et al. (2020). Expression atlas update: from tissues to single cells. *Nucleic Acids Res.* 48, D77–83. doi: 10.1093/nar/gkz947

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., et al. (2023). InterPro in 2022. *Nucleic Acids Res.* 51, D418–D427. doi: 10.1093/nar/gkac993

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* 47, D1146–D1154. doi: 10.1093/nar/gky1046

Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., et al. (2021). KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response. *Patterns.* 2, 100155. doi: 10.1016/j.patter.2020.100155

Reiser, L., Subramaniam, S., Li, D., and Huala, E. (2022). Using the Arabidopsis Information Resource (TAIR) to find information about arabidopsis genes. *Current Protocols.* 2, e574. doi: 10.1002/cpz1.574

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice annotation project database (rap-db): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54, e6. doi: 10.1093/pcp/pcs183

Sheth, A., Padhee, S., and Gyrard, A. (2019). Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Comput.* 23, 67–75. doi: 10.1109/MIC.2019.2928449

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., and Hancock, D. (2010). Multifunctional crop trait ontology for breeders' data: Field book, annotation, data discovery and semantic enrichment of the literature. *AoB Plants.* 2010, lq008. doi: 10.1093/aobpla/plq008

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (Fragaria Vesca). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Silverman, E. K., Schmidt, H. H. H. W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., et al. (2020). Molecular networks in Network Medicine: Development and applications. *Syst. Biol. Med.* 12, e1489. doi: 10.1002/wsbm.1489

Spooner, W., Youens-Clark, K., Staines, D., and Ware, D. (2012). GrameneMart: The BioMart data portal for the gramene project. *Datab.* 2012, bar056. doi: 10.1093/database/bar056

Tello-Ruiz, M. K., Naithani, S., Stein, J. C., Gupta, P., Campbell, M., Olson, A., et al. (2018). Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.* 46, D1181–D1189. doi: 10.1093/nar/gkx1111

The GO Consortium (2021). The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.* 49, D325–D334. doi: 10.1093/nar/gkaa1113

Thudi, M., Palakurthi, R., Schnable, J. C., Chitikineni, A., Dreisigacker, S., Mace, E., et al. (2021). Genomic resources in plant breeding for sustainable agriculture. *J. Plant Physiol.* 257, 153351. doi: 10.1016/j.jplph.2020.153351

Tian, Z., Wang, J. W. L., i,. J., and Han, B. (2021). Designing future crops: challenges and strategies for sustainable agriculture. *Plant J.* 105, 1165–1178. doi: 10.1111/tpj.15107

Tiddi, I., and Balliet, D., ten Teije, A. (2020). "Fostering scientific meta-analysis with knowledge graphs: A case study," in *The Semantic Web ESWC 2020. Lecture Notes in Computer Science* (Cham: Springer) 287–303. doi: 10.1007/978-3-030-49461-2_17

Tirosh, I., Weinberger, A., Bezalel, D., Kaganovich, M., and Barkai, N. (2008). On the relation between promoter divergence and gene expression evolution. *Mol. Syst. Biol.* 4, 159. doi: 10.1038/msb4100198

Unni, D. R., Moxon, S. A. T., Bada, M., Brush, M., Bruskiewich, R., Caufield, J. H., et al. (2022). Biolink model: a universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clin. Transl. Sci.* 15, 1848–1888. doi: 10.1111/cts.13302

Walls, R. L., Cooper, L., Elser, J., Gandolfo, M. A., Mungall, C. J., Smith, B., et al. (2019). The plant ontology facilitates comparisons of plant development stages across species. *Front. Plant Sci.* 10, 631. doi: 10.3389/fpls.2019.00631

Zemojtel, T., Köhler, S., Mackenroth, L., Jäger, M., Hecht, J., Krawitz, P., et al. (2014). Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci. Transl. Med.* 6, 252ra123. doi: 10.1126/scitranslmed.3009262

Appendix

All supplementary files can be accessed in GitHub under a CC-0 license (https://github.com/diatomsRcool/supplementary_material/tree/main/promoter_region).

1. drought_expression.tsv
2. drought_genes.tsv
3. GO_annotations.tsv
4. panther_results folder
5. promoter_region_clustvis_data0.tsv
6. orthologous_genes.tsv

Clustvis analysis is at https://biit.cs.ut.ee/clustvis/?s=IWJNurmUtGWZoMt.

TYPE Original Research
PUBLISHED 19 June 2023
DOI 10.3389/frai.2023.1131667

EPPO ontology: a semantic-driven approach for plant and pest codes representation

Aarón Ayllón-Benitez[1]*, José Antonio Bernabé-Diaz[1], Paola Espinoza-Arias[1], Iker Esnaola-Gonzalez[1], Delphine S. A. Beeckman[2], Bonnie McCaig[3], Kristin Hanzlik[4], Toon Cools[5], Carlos Castro Iragorri[6] and Nicolás Palacios[6]

[1]BASF Digital Solutions, Madrid, Spain, [2]BASF Belgium Coordination Center CommV, Innovation Center Gent, Ghent, Belgium, [3]BASF Corporation, Raleigh, NC, United States, [4]BASF SE Data Management and Data Governance, Global Research Services APR/HP, Limburgerhof, Germany, [5]TalentBay, Brussels, Belgium, [6]Linking Data SAS, Bogotá, Colombia

The agricultural industry and regulatory organizations define strategies and build tools and products for plant protection against pests. To identify different plants and their related pests and avoid inconsistencies between such organizations, an agreed and shared classification is necessary. In this regard, the European and Mediterranean Plant Protection Organization (EPPO) has been working on defining and maintaining a harmonized coding system (EPPO codes). EPPO codes are an easy way of referring to a specific organism by means of short 5 or 6 letter codes instead of long scientific names or ambiguous common names. EPPO codes are freely available in different formats through the EPPO Global Database platform and are implemented as a worldwide standard and used among scientists and experts in both industry and regulatory organizations. One of the large companies that adopted such codes is BASF, which uses them mainly in research and development to build their crop protection and seeds products. However, extracting the information is limited by fixed API calls or files that require additional processing steps. Facing these issues makes it difficult to use the available information flexibly, infer new data connections, or enrich it with external data sources. To overcome such limitations, BASF has developed an internal EPPO ontology to represent the list of codes provided by the EPPO Global Database as well as the regulatory categorization and relationship among them. This paper presents the development process of this ontology along with its enrichment process, which allows the reuse of relevant information available in an external knowledge source such as the NCBI Taxon. In addition, this paper describes the use and adoption of the EPPO ontology within the BASF's Agricultural Solutions division and the lessons learned during this work.

KEYWORDS

EPPO codes, ontologies, plants, seeds, diseases, pests, crop protection, chemical industry

Check for updates

OPEN ACCESS

EDITED BY
Catherine Roussey,
INRAE Occitanie Montpellier, France

REVIEWED BY
Claire Nédellec,
INRAE Centre Ile de France, France
Nadia Yacoubi Ayadi,
Research Centre Inria Sophia Antipolis Méditerranée, France
Jean Charlet,
Assistance Publique Hopitaux De Paris, France

*CORRESPONDENCE
Aarón Ayllón-Benitez
✉ aaron.ayllon@basf.com

RECEIVED 25 December 2022
ACCEPTED 26 May 2023
PUBLISHED 19 June 2023

CITATION
Ayllón-Benitez A, Bernabé-Diaz JA, Espinoza-Arias P, Esnaola-Gonzalez I, Beeckman DSA, McCaig B, Hanzlik K, Cools T, Castro Iragorri C and Palacios N (2023) EPPO ontology: a semantic-driven approach for plant and pest codes representation. *Front. Artif. Intell.* 6:1131667. doi: 10.3389/frai.2023.1131667

1. Introduction

In agriculture, reducing crop losses caused by organisms such as pests and diseases is crucial. In 2021 it was estimated that up to 40 percent of global crop production is lost annually due to pests (IPPC Secretariat et al., 2021), leading to huge economic costs, low availability and quality of food and raw materials, and environmental pollution, among others negative effects. In the last decades several organizations and companies have been working to provide regulations, technologies and products to prevent and mitigate damage caused by pests outbreaks. Therefore, to have a common and consistent way of identifying plants and pests when providing their solutions, such organizations and companies use the EPPO coding system as the worldwide reference.

The EPPO coding system was created and maintained by Bayer in the 1970s and then transferred to the European and Mediterranean Plant Protection Organization (EPPO) in 1996. In 2014, this system was released as the EPPO Global Database,[1] freely available under an open data license and in several formats (e.g., XML, SQLite, TXT). In this coding system, an EPPO code is a unique identifier for plants, pests, and pathogens which is built as combinations of 5 to 6 letters. EPPO codes mainly cover taxonomic codes but also non-taxonomic codes. On the one hand, taxonomic codes refer to those EPPO codes developed for biological organisms or groups of biological organisms based on their scientific naming and classification in groups known as "taxa". On the other hand, non-taxonomic codes represent a smaller set of codes describing entities of interest to those working in the field of plant protection products (PPP). Developed with the aim to describe the use of a PPP, they facilitate communication among National Plant Protection Organizations and other stakeholders involved in the registration of plant protection products. Further details on the information available for taxonomic and non-taxonomic codes are given in Figure 1.

In addition, EPPO codes are hierarchically organized and, specifically within the taxonomic portion of the EPPO Global Database, each taxonomic level has a unique code which is mainly derived from the corresponding scientific name of that level. Whereas, in the case of non-taxonomic codes, they are built following more concrete rules described in the EPPO Standard PP1/248 (European and Mediterranean Plant Protection Organization, 2022). Currently, the EPPO database contains basic information of more than 90,000 species and detailed information for more than 1,700 pests and diseases. Even so, the coding system is dynamic and new codes constantly are added [on average more than 2,000 new codes per year (Roy, 2019)].

BASF is one of the large companies consuming EPPO codes as a standard for plant pest identification. BASF applies EPPO codes in the research and development of new agricultural products (such as insecticides, fungicides, herbicides, seeds, among others) and tools (e.g., a system for disease and pest recognition, and tailored recommendation of treatments based on in-field conditions[2]). Nevertheless, the availability of multiple format files to extract EPPO codes data requires additional processing steps to consume them. To solve this, EPPO Global Database provides a fixed REST API to extract data; however, this limits the flexibility of data consumption. Therefore, consuming the information of EPPO codes requires accessing to these different files and API requests to get the complete information needed. To face these limitations and provide more capabilities to EPPO codes, we developed an ontology to represent them in a formal semantic language. Ontologies allow to homogeneously structure and harmonize data without ambiguities, infer new knowledge, and enrich data with external knowledge sources (Studer et al., 1998). The adoption of ontologies in large companies like BASF allows sharing and reusing common parts of knowledge across the organization, facilitating data reusability and interoperability.

In this manuscript, we detail the process followed to build the EPPO ontology and the lessons learned during this work. We begin by describing the related work (Section 2). Then, we explain the ontology development process along with its automatic creation pipeline and the enrichment step (Section 3). Next, we describe in detail the main ontology elements (Section 4) and illustrate how BASF is using the EPPO ontology (Section 5). Finally, we outline our conclusions and discuss future work (Section 6).

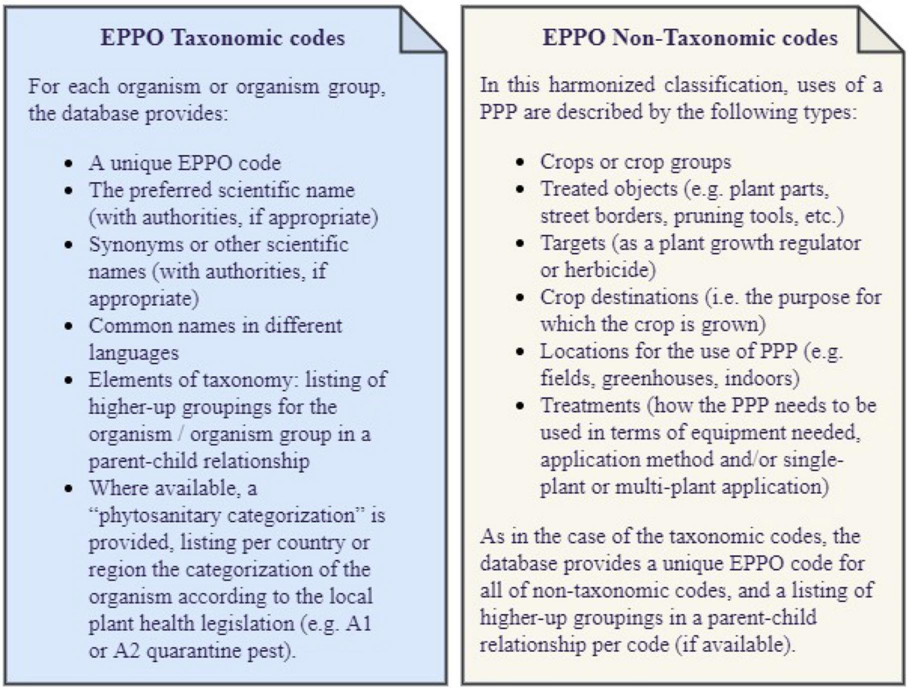
2. Related work

In the context of this work, some ontologies have been reported in the literature. From a general point of view, the most relevant ontology for our work is the NCBITaxon ontology (Bastian et al., 2013) which allows describing organism names and taxonomic lineages from the NCBI taxonomy database (Federhen, 2012). This ontology provides a comprehensive collection of organisms including the taxonomic levels (e.g., kingdom, order, family, etc.) that are also detailed in the EPPO codes. However, it does not include further information, provided by the EPPO Global Database representation such as code, EPPO code phytosanitary categorization, categorization status, code type, host-pests relationship, etc.

Focusing on plant pests and diseases, few ontologies have been reported to represent the crop domain including pests. The Pest Crop Ontology (PCO) (Damos et al., 2017) provides a high-level representation of crops, pests, treatments, and the relations among them. To provide further details than those provided by PCO, the Pests in Crops and their Treatments Ontology (PCT-O) (Lacasta et al., 2018) was developed to describe the conditions required by a pest to produce outbreaks and the restrictions on the treatments. In terms of describing crop management details, the Crop Planning and Production Process Ontology (C3PO) (Darnala et al., 2021) allows representing plot management and crop itineraries by means of several modules which encapsulate high-level information about plants, crop management, potential diseases and pests, treatments, among others. However, none of the aforementioned ontologies include further details on pests such as a consistent taxon, non-taxon, and commodity group classification, synonyms, preferred names, and granular details about them. Finally, the Plant Health Threat Ontology (Alomar et al., 2015, 2016) formally represents plant pest and disease names and the relations among them and to other concepts like hosts, symptoms, crops, etc. This ontology reuses the Plant Ontology (Cooper et al., 2013), and concepts

[1] https://gd.eppo.int

[2] https://www.xarvio.com/global/en.html



EPPO Taxonomic codes

For each organism or organism group, the database provides:

- A unique EPPO code
- The preferred scientific name (with authorities, if appropriate)
- Synonyms or other scientific names (with authorities, if appropriate)
- Common names in different languages
- Elements of taxonomy: listing of higher-up groupings for the organism / organism group in a parent-child relationship
- Where available, a "phytosanitary categorization" is provided, listing per country or region the categorization of the organism according to the local plant health legislation (e.g. A1 or A2 quarantine pest).

EPPO Non-Taxonomic codes

In this harmonized classification, uses of a PPP are described by the following types:

- Crops or crop groups
- Treated objects (e.g. plant parts, street borders, pruning tools, etc.)
- Targets (as a plant growth regulator or herbicide)
- Crop destinations (i.e. the purpose for which the crop is grown)
- Locations for the use of PPP (e.g. fields, greenhouses, indoors)
- Treatments (how the PPP needs to be used in terms of equipment needed, application method and/or single-plant or multi-plant application)

As in the case of the taxonomic codes, the database provides a unique EPPO code for all of non-taxonomic codes, and a listing of higher-up groupings in a parent-child relationship per code (if available).

FIGURE 1
Taxonomic vs. non-taxonomic codes.

coming from multilingual sources such as UniProt Taxon, EPPO Global Database and DBPedia. In terms of EPPO information, a recent report (European Food Safety Authority et al., 2021) details that 133 plant pests are included in the current ontology version. Unfortunately, this ontology is not publicly available; therefore, it is not possible to analyze it and, consequently, the plant pests that it represents cannot be reused.

3. Development of the EPPO ontology

The ontology was built following the development lifecycle proposed in the BASF Governance Operational Model for Ontologies (GOMO) (Iglesias-Molina et al., 2022). This lifecycle was derived from the Linked Open Terms (LOT) methodology (Poveda-Villalón et al., 2022), which is a methodology based on agile techniques and comprises several stages and activities for the ontology construction. The GOMO lifecycle includes four main stages which will be described in the following subsections.

3.1. Requirements and kick off

This stage intends to define and gather all the requirements and basic elements necessary for the ontology development. Therefore, the first activity we undertook was to define the purpose and scope of the EPPO ontology. To do so, we collected the feedback of several domain experts from our Agricultural Solutions division and agreed that the purpose covered by this ontology is the representation of the information available in the EPPO Global database and the relationships between the concepts identified therein. Therefore, this ontology is not limited to be used by a specific application, but has been developed in the interest of having a single, harmonized, and flexible source of the EPPO code system information. As for the ontology scope, we agreed to include taxonomic and non-taxonomic codes along with their code types, parent-child relationship per code, phytosanitary categorization, and their taxonomy level (if applicable). Further details on the information available in taxonomic and non-taxonomic classifications is presented in Figure 1.

The second activity we performed was to define the requirements that the ontology must fulfil. To this end and based on the needs of the domain experts, we posed several competency questions (Grüninger and Fox, 1995) that guided us during the development process. Table 1 shows an excerpt of the competency questions. A complete list is provided in Section 1 of the Supplementary material.

The third activity we executed was to identify and analyze the structure of the relevant data sources in relation to the ontology purpose and scope. We identified several files in the EPPO Data Services; however, we focused particularly on three of them:

(a) the SQLite database file[3] containing EPPO codes for taxonomic and non-taxonomic organisms, including data such as their preferred names, synonyms in several languages, creation and modification dates, among others; (b) the REST API service[4] that provides direct access to information specific to EPPO codes,

[3] https://data.eppo.int/files/sqlite_all.zip

[4] https://data.eppo.int/documentation/rest#collapse1

TABLE 1 Excerpt of competency questions from the EPPO ontology.

Identifier	Competency question	Expected answer
CQ1	Which taxonomic code is associated with non-taxonomic code "TRZAW"?	https://ontology.basf.net/ontology/BASF/Bioscience/EPPO/TRZAX
CQ2	List the non-taxonomic EPPO codes + names associated with Species ("Brassica juncea") or EPPO Code ("BRSJU") (Is this species part of any crop group?)	non-taxonomic EPPO code: https://ontology.basf.net/ontology/BASF/Bioscience/EPPO/BRSJU, non-taxonomic EPPO name: `leafy brassica crops`; non-taxonomic EPPO code: https://ontology.basf.net/ontology/BASF/Bioscience/EPPO/3MUSC, non-taxonomic EPPO name: `mustard crops`
CQ3	Do "BRSJU" and "BRSRW" belong to a common crop group?—leafy brassica crops (3LFBC)	True
CQ4	List all EPPO Codes (+ names + description) that are part of non-taxonomic code group "treatment methods" (3TMETM)	EPPO code: https://ontology.basf.net/ontology/BASF/Bioscience/EPPO/3BRUSM, EPPO name: `brushing`, EPPO description: `Application of a liquid product or powder with a brush, e.g., tree trunk application of fungicide in citrus or local treatment of single weeds in a crop stand`; ...

e.g., to their taxonomy classification, categorization list, hosts, pests, among others; (c) the Replaced codes[5] file, which contains information on the entire history of EPPO codes that were superseded by other EPPO codes. Finally, we also took into account several so-called "categorization" lists,[6] available in the EPPO Global Database web page. These lists indicate what the regulatory status from a phytosanitary (i.e., plant health) perspective is for a given organism (EPPO code) as defined by a Regional Plant Protection Organization (RPPO), based on the local plant health legislation (e.g., A1 or A2 quarantine pest).

Lastly, in the fourth activity we identified a reusable terminology resource relevant to the ontology purpose and scope. More specifically, we chose the NCBITaxon ontology[7] (explained in Section 2) as the most related resource to be reused during the ontology enrichment activity.

3.2. Implementation

This stage aims to generate the ontology based on the requirements and data sources previously identified. For this purpose, the first activity we carried out was to build a conceptual model to define the classes and properties that represent the ontology domain. We defined such model as a diagram following the details of the Chowlk notation (Chávez-Feria et al., 2022), which is a UML-based notation for ontology diagrams. Figure 2 shows the conceptualization diagram we defined for the EPPO ontology. Note that, due to the large number of terms contained in the EPPO Global database, this diagram only shows the main classes and properties represented in the ontology. However, the ontology contains all the hierarchical classifications included in the database for each class depicted in the diagram.

Next, taking as input the structure we defined in the conceptual model, the second activity we performed was the ontology encoding. The goal of this activity was to generate the ontology as a machine-readable model in an ontology representation language. Figure 3 depicts the steps we carried out to generate the ontology. First, we performed a transformation of non-ontological resources (the data sources identified in the previous stage) into an ontological one. This transformation task was mainly performed automatically using a Python package (eppo_tools) that we implemented for this purpose. This package reuses pre-existing and well-know libraries such as Requests,[8] SQLAlchemy,[9] lxml,[10] RDFLib,[11] among others that allow us to access the data sources, manage the data and build the ontology code. As a result we obtained the ontology encoded in the Web Ontology Language (OWL). Then, as the different types of EPPO phytosanitary categorizations were extracted from the EPPO Global database web page, human intervention was needed to define such categories and their taxonomy in the ontology. For the human intervention, a domain expert lead the manual extraction and definition of the categorization lists in the ontology using the WebProtégé ontology editor (Tudorache et al., 2013). It is worth mentioning that we also use such editor to add relevant ontology metadata (e.g., creator, title, license, among others) which is useful for ontology reusability purposes. Finally, it is important to note that the EPPO ontology reuses several properties from other ontologies. To this end, we applied the *soft reuse* technique which allows referencing the reused ontology elements URIs instead of importing the whole ontology (*hard reuse*) (Fernandez-Lopez et al., 2019). To decide which properties to reuse, we first analyze the semantics of each property and also look at how common its use is in the community.

Then, the third activity we conducted was the ontology enrichment, which is also depicted in Figure 3. The main objective of such activity was to automatically map NCBITaxon IRIs to EPPO ontology elements that match specific annotations (e.g., `rdfs:label`[12] or `skos:altlabel`[13]).[14] To generate

[5] https://data.eppo.int/files/replaced.zip

[6] https://gd.eppo.int/rppo/

[7] https://obofoundry.org/ontology/ncbitaxon.html

[8] https://pypi.org/project/requests/

[9] https://www.sqlalchemy.org/

[10] https://pypi.org/project/lxml/

[11] https://rdflib.readthedocs.io/

[12] http://www.w3.org/2000/01/rdf-schema#label

[13] http://www.w3.org/2004/02/skos/core#altLabel

[14] Note that in the manuscript we use typewriter font when referring to parts of the ontology code. In addition, when reusing elements from another ontology, their prefix is included before the colon and then its local identifier is included.

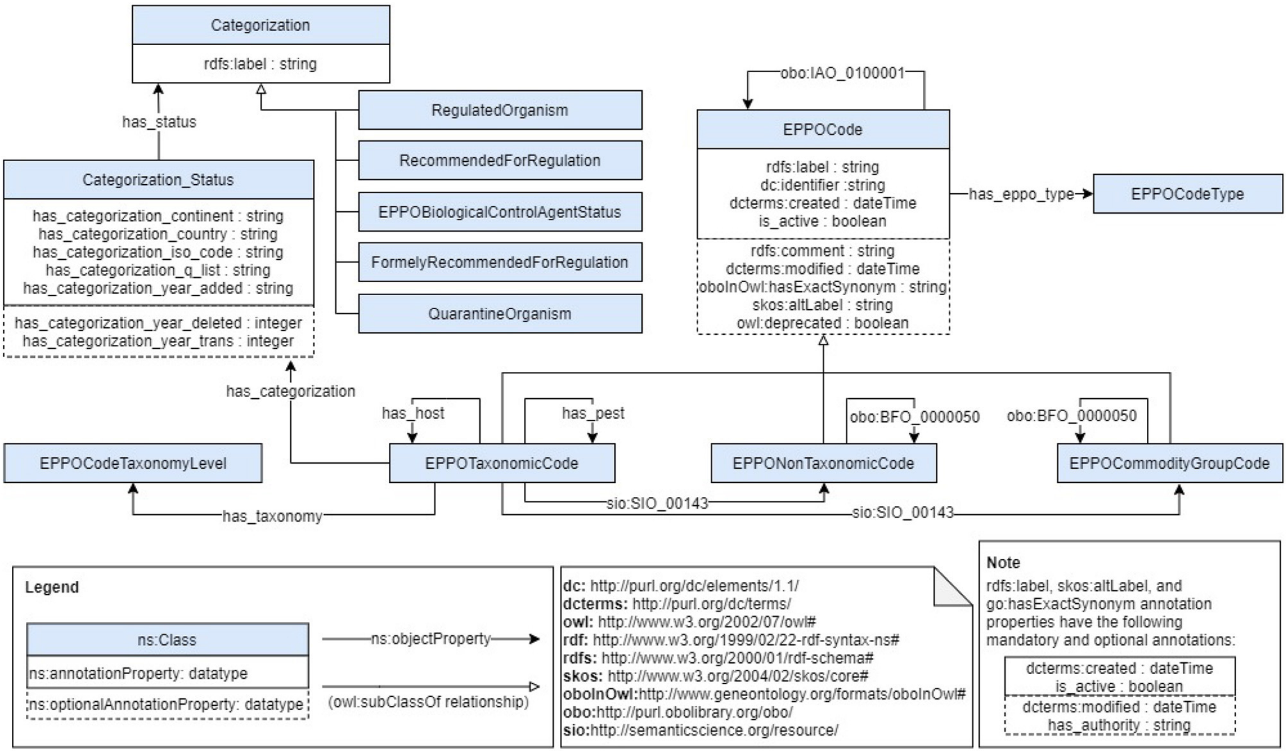

Categorization
rdfs:label : string
has_status
Categorization_Status
has_categorization_continent : string
has_categorization_country : string
has_categorization_iso_code : string
has_categorization_q_list : string
has_categorization_year_added : string
has_categorization_year_deleted : integer
has_categorization_year_trans : integer
RegulatedOrganism
RecommendedForRegulation
EPPOBiologicalControlAgentStatus
FormelyRecommendedForRegulation
QuarantineOrganism
obo:IAO_0100001
EPPOCode
rdfs:label : string
dc:identifier :string
dcterms:created : dateTime
is_active : boolean
rdfs:comment : string
dcterms:modified : dateTime
oboInOwl:hasExactSynonym : string
skos:altLabel : string
owl:deprecated : boolean
has_eppo_type
EPPOCodeType
has_categorization
has_host
has_pest
obo:BFO_0000050
obo:BFO_0000050
EPPOCodeTaxonomyLevel
EPPOTaxonomicCode
EPPONonTaxonomicCode
EPPOCommodityGroupCode
has_taxonomy
sio:SIO_00143
sio:SIO_00143
Legend
ns:Class
ns:annotationProperty: datatype
ns:optionalAnnotationProperty: datatype
ns:objectProperty
(owl:subClassOf relationship)
dc: http://purl.org/dc/elements/1.1/
dcterms: http://purl.org/dc/terms/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
skos: http://www.w3.org/2004/02/skos/core#
oboInOwl:http://www.geneontology.org/formats/oboInOwl#
obo:http://purl.obolibrary.org/obo/
sio:http://semanticscience.org/resource/
Note
rdfs:label, skos:altLabel, and go:hasExactSynonym annotation properties have the following mandatory and optional annotations:
dcterms:created : dateTime
is_active : boolean
dcterms:modified : dateTime
has_authority : string

FIGURE 2
EPPO ontology conceptualization diagram.

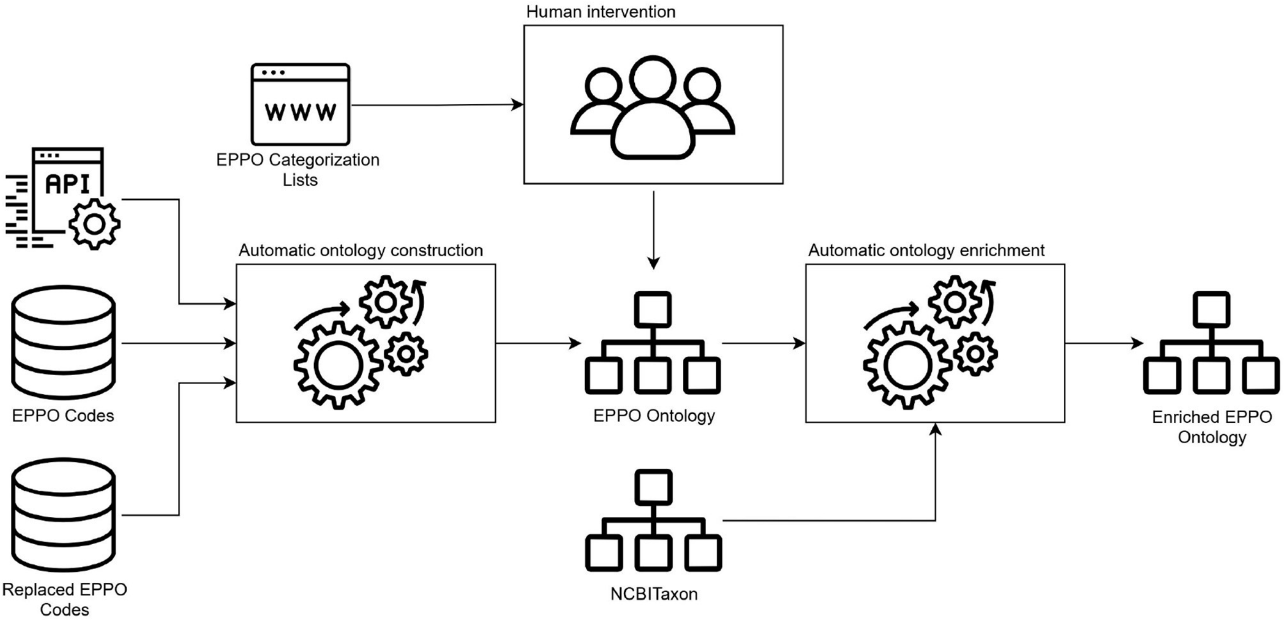


FIGURE 3
Ontology encoding and enrichment pipeline.

such mappings we built Python scripts to automatically include the NCBITaxon IRIs into the EPPO ontology. It is worth mentioning that to include the mappings we also reused ROBOT (Jackson et al., 2019), which is an open-source library and command-line tool to automate ontology development tasks. The mappings were included in the EPPO ontology by means of `oboInOwl:hasDbXref`[15] property which represents a reference to an identical or very similar object in another resource. As a result of this activity we obtained an enriched EPPO ontology. More details of the mapping process are provided in Section 3 of the Supplementary material.

Finally, in the fourth activity we evaluated the ontology to verify that it was correctly built according to the competency questions formulated in the Requirements/Kick off stage. To do this, we translated the competency questions into SPARQL queries in order to run them against the ontology to obtain the expected answers. The SPARQL queries we generated for the ontology evaluation are provided in Section 2 of the Supplementary material.

3.3. Publication

This stage aims to deliver the ontology online as human-readable documentation and as a machine-readable file. As for the documentation, we built an HTML file to include a human-readable description of the ontology design that includes diagrams and details about the main the classes and properties. In addition, it includes guidelines on the Python package that bundles all the functionality related to the automatic ontology generation. This HTML documentation is published internally and is made available in the BASF intranet. Finally, to facilitate searching and browsing of the ontology, it is registered in our internal Ontology Lookup Service (OLS) (Côté et al., 2006). This service provides a user-friendly interface with search mechanisms that makes the ontology findable by anyone in the company. OLS also makes use of the ontology metadata to display it to users so that they can analyze the ontology in detail. The latest version of the EPPO ontology is available in our BASF GitHub repository.[16]

3.4. Maintenance

Ontologies may degrade over time, due to different reasons including changes or additions in the domains the ontology is modeling, a changing view of the world or a change in usage perspective (Noy and Musen, 2003; Tartir et al., 2010). Therefore, a methodical approach to handle, manage and adapt to changes is of utmost importance during an ontology lifecycle. In our case, as we mentioned in Section 1, EPPO codes are not a static data source; therefore, codes can change or new ones may be added. Such a dynamic environment requires a well-defined strategy to ensure that users have access to the latest available knowledge, and this strategy consists of an automated run of our Python package whenever the public EPPO SQLite file is updated. Then, if a new categorization list appears in the latest version of the database, we inform our domain experts so that they can manually classify it in the corresponding class or, if necessary, create a new class in which to classify it. Lastly, our mappings to the NCBITaxon are also run to ensure that the new version of the ontology contains the references to that external knowledge source.

TABLE 2 EPPO ontology metrics taken from the Protégé ontology editor.

Ontology elements	Count
Axioms	2,191,211
Logical axioms	49,2712
Declaration axioms	13,8149
Classes	13,8099
Object properties	20
Annotation properties	35

As defined in our GOMO best practices, the maintenance process is performed in a git repository, where we use different environments to deal with ontology changes. Whenever, an update to the ontology occurs, it is deployed to the *DEV* environment, which contains work in progress not yet available to end users. Likewise, the content is also deployed into the QA (Quality Assessment) environment, where users can access and notify potential problems they may encounter in the updated version. Then, once a week, the ontology from *QA* is deployed to the *PROD* environment, which involves the ontology release. New ontology releases are notified to EPPO ontology users through our internal communication channels, so that they are well-aware of the new information available.

4. EPPO ontology description

In this section, we provide further details on the ontology in terms of its main metrics and structure. First, we present the ontology metrics which are listed in Table 2. Such table presents the count of the different ontology elements we generated. In summary, we created more than 130 thousand classes, 20 object properties, and 35 annotation properties which allow representing the EPPO codes concepts, their attributes and the relationships among such concepts.

Then, we provide further details on the ontology structure that was previously depicted in our conceptualization model shown in Figure 2. Note that all prefixes used in this section are listed in Figure 2. The following subsections describe the most relevant classes and properties of the ontology as well as the main relationships among such classes. Finally, we present an example of the ontological representation of an EPPO code using an excerpt from the EPPO ontology.

[15] http://www.geneontology.org/formats/oboInOwl#hasDbXref

[16] https://github.com/basf/EPPOontology

4.1. EPPO code

It represents the core class of our ontology, as it contains the most relevant information about codes and their links to all the EPPO Code names. In addition, it is the parent class of several concepts, such as Taxonomic, Non-Taxonomic, and Commodity Group, which allow representing the codes in a more granular way. As previously described in this work, Taxonomic codes represent organisms or organisms groups known as taxa, and Non-Taxonomic codes represents entities of interest for PPP. As for Commodity Group codes, they represent a subset of codes which allow grouping plant commodities (e.g., fruit plants, aquarium plants, conifers, etc.) liable to spread a pest in international trade.

Going into more details of the EPPO Code, each code contains information about its name (`rdfs:label`), creation date (`dcterms:created`[17]), and whether it is active or not (`isActive`). Optionally, a code can also contain a synonym (`oboInOwl:hasExactSynonym`[18]), alternative name (`skos:altLabel`), modification date (`dcterms:modified`[19]), whether it is deprecated or not (`owl:deprecated`[20]), and definition (`rdfs:comment`[21]). More precise details are also included in the name, alternative name, and synonym properties, since the ontology also represents their creation and modification dates, whether they are active or not, and what was the Name's authority (`has_authority`), e.g., *Gennadius*. It is worth mentioning that all names and synonyms have a corresponding language tag. As for the scientific name (preferred name) and other scientific names the language tag assigned is Latin (la), since it is the official language in which scientific names are defined and, therefore, the language provided by the database. While for common names the language tag is assigned depending on the language in which it is available in the database.

Furthermore, many EPPO Codes belonging to the Taxonomic Code class include information about their phytosanitary status, which represents the categorization list in which they have been classified. For this purpose, such codes are linked to their corresponding Categorization Status by means of the `has_categorization` property. In addition, EPPO Codes represent their corresponding taxonomy level (`has_taxonomy_level`), i.e., the integer value representing the distance between a term and its higher-level taxonomic group. Finally, these codes can also include information about their hosts or pests (`has_host` or `has_pest`) to represent host-pests or pest-hosts relationship. In regard to the `has_pest` property, it holds several subproperties which represent all the categories of pest/host plant combinations provided in the database.[22] For example, the "Alternate" category is represented by the `has_pest_type_alternate` subproperty which defines a relationship between an organism and the distinct hosts it needs to complete its life cycle. As for the `has_host` property, it represents the inverse property of `has_pest` property. For example, the `has_host_type_alternate` subproperty represents a host which is used by a pest during its life cycle.

Finally, to represent specific relationships among codes, the ontology reuses two properties: (1) the `obo:BFO_0000050`[23] property to represent that a Non-Taxonomic or Commodity Group code is part of a subset of them, and (2) the `sio:SIO_001403`[24] property to represent that a Taxonomic or a Non-Taxonomic code is associated with a Commodity Group code.

4.2. Categorization status

This class contains phytosanitary categorization for a given EPPO Code in a region or country, based on the corresponding specific RPPO phytosanitary categorization list (`has_status`) and a nomenclature for that list as defined in the EPPO Global Database (`categorization_q_list`, note that "q" stands for "quarantine"). To provide granular details of a categorization, it includes the continent (`has_categorization_continent`) and country (`has_categorization_country`) names, and the ISO country code (`has_categorization_iso_code`) to which the list is applicable. Relevant dates are also represented for each categorization list, such as the year it was added (`has_categorization_year_added`), the year it was removed (`has_categorization_year_deleted`) or the year it was transferred (`has_categorization_year_trans`) to another categorization.

4.3. Categorization

This class represents the general types of categorizations in which EPPO Codes may be listed. These categorizations are used to draw the attention of countries and regions to the status of plant pests and diseases in terms of the potential phytosanitary risks that they may pose. For example, a pest categorized as part of a quarantine list (`QuarantinePest`) constitutes a regulatory requirement in terms of phytosanitary measures to be implemented for that pest. As mentioned in Section 3.2, the Code Categorization class contains a hierarchy manually generated by our domain experts. This hierarchy provides a higher-up grouping for the categorizations existing in the EPPO Global database. For example, a quarantine list (`QuarantinePest`) belongs to (`rdfs:subclassOf`) the quarantine organism (`QuarantineOrganism`) class defined

[17] http://purl.org/dc/terms/created

[18] http://www.geneontology.org/formats/oboInOwl#hasExactSynonym

[19] http://purl.org/dc/terms/modified

[20] http://www.w3.org/2002/07/owl#deprecated

[21] http://www.w3.org/2000/01/rdf-schema#comment

[22] Further explanation of categories is available in the Host Plants section of the EPPO Global Database guide: https://gd.eppo.int/media/files/general_user-guide.pdf.

[23] http://purl.obolibrary.org/obo/BFO_0000050

[24] https://semanticscience.org/resource/SIO_001403

by our experts. More details on the definition of the Code Categorization hierarchy are provided in Section 4 of the Supplementary material.

4.4. EPPO code taxonomy level

This class defines the different types of taxonomy levels, such as Kingdom, Family, or Species, among others, to which an EPPO Code belongs. To this end, each code is related to its taxonomic level by means of the `has_taxonomy` property. It is worth mentioning that only those codes that belong to the Taxonomic Code class can be linked to a taxonomy level. Finally, each taxonomy level contains a cross-reference to its corresponding term defined in the NCBITaxon ontology.

4.5. EPPO code type

This class allows representing a more granular classification of the EPPO codes to group them into different levels: species level, higher taxonomic group of organisms, or non-taxonomic entities. For taxonomic EPPO codes at species level, the EPPO Code Type class distinguishes between plant, animal, and microorganism. As for higher taxonomic groups (e.g., genus, family etc.) it includes plant taxonomic group, animal taxonomic group, and microorganism taxonomic group. For other non-taxonomic entities it includes non-taxonomic and commodity groups. In addition to its label, each type also contains the identifier assigned by the coding system. Finally, EPPO Codes are related to their specific code type by means of the `has_eppo_type` property.

4.6. EPPO replaced codes

As mentioned earlier during the ontology development process, the ontology also represents the superseded codes available in the EPPO Global Database. To this end, all these codes contain similar properties to those included in the EPPO Codes that are still active. However, the Replaced codes have two annotation properties that allow them to be identified as part of the coding system archive. First, the boolean property defined in the ontology to represent whether a code is active (`isActive`) is declared as false. Second, following our GOMO Standard for deprecation of ontology elements, the boolean property defined to specify that an IRI is deprecated (`owl:deprecated`) is declared as true. In addition, the term that replaces the code is defined with the `obo:IAO_0100001`[25] property that allows the term to be related. to another term that is used as a substitute. In this manner, the EPPO ontology also represents codes that are not active but that can be relevant for traceability purposes.

4.7. Example of the ontology representation of an EPPO code

In order to illustrate how the main classes and properties have been defined in the ontology, we present an example that represents the information of an EPPO code using the ontology elements. For this purpose, we use the information from the TRZAW code (which is the code referred to in the first Competency Question presented in Table 1). The most relevant information of this code can be retrieved from the "Overview" menu of the EPPO Global Database website, as shown in Figure 4. As can be seen in this figure, the TRZAW code is presented as a non-taxonomic code, along with its code, preferred scientific name, and other common names in different languages. In addition, a classification tree is presented to navigate through the hierarchy to which it belongs. Moreover, TRZAX is shown as the taxon associated to the TRZAW code (note that this relationship provides the answer to our first Competency Question). Finally, the creation date of the code is also shown.

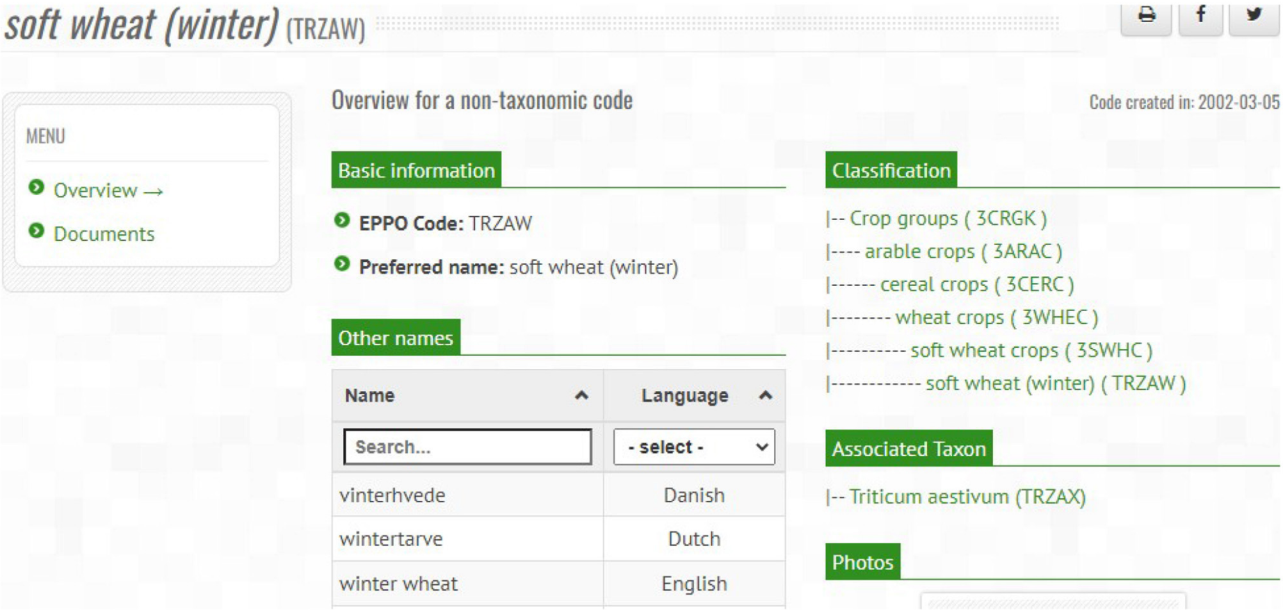
The ontological representation of the information shown above for the TRZAW code is provided in Listing 1. This listing (written in Turtle[26] format) is an excerpt from the EPPO ontology that also includes extra information that is not retrieved from the TRZAW code overview presented in Figure 4. Going into detail, this listing begins with the definition of the TRZAW code as a subclass of the `NonTaxonomicCode` class and its linkage to the 3SWHC code (*soft wheat crops*[27]) via the part of (`obo:BFO_0000050`) property. In addition, several properties have been defined to represent the values of TRZAW's preferred name (`rdfs:label`), other name (`skos:altLabel`), EPPO code (`dc:identifier`), creation and modification dates (`dcterms:created` and `dcterms:modified`), other common names in different languages (`hasExactSynonym`),[28] active status (`is_active`), and its specific code type (`has_eppo_type`). The TRZAW code type corresponds to Non Taxonomic (`NTX`), which is defined later in this listing as a subclass of the `EPPOCodeType` class along with its code (`dc:identifier`), and name (`rdfs:label`). Moreover, the TRZAW code contains a reference to a similar term of the NCBITaxon. This reference is represented by the `oboInOwl:hasDbXref` property and its value corresponds to *Triticum aestivum* (`obo:NCBITaxon_4565`). Finally, it should be noted that `rdfs:label`, `skos:altLabel`, and `oboInOwl:hasExactSynonym` properties contain additional annotations (`dcterms:created` and `is_active`),

[25] http://purl.obolibrary.org/obo/IAO_0100001

[26] https://www.w3.org/TR/turtle

[27] Note that the 3SWHC code definition is not included in this listing, but is represented in the ontology using similar properties and structure as the TRZAW code presented in this example.

[28] Note that, for simplicity, we have included few synonyms for the codes shown in this listing.



soft wheat (winter) (TRZAW)

MENU

- Overview →
- Documents

Overview for a non-taxonomic code

Code created in: 2002-03-05

Basic information

- **EPPO Code:** TRZAW
- **Preferred name:** soft wheat (winter)

Other names

Name	Language
Search...	- select -
vinterhvede	Danish
wintertarve	Dutch
winter wheat	English

Classification

|-- Crop groups (3CRGK)
|---- arable crops (3ARAC)
|------ cereal crops (3CERC)
|-------- wheat crops (3WHEC)
|---------- soft wheat crops (3SWHC)
|------------ soft wheat (winter) (TRZAW)

Associated Taxon

|-- Triticum aestivum (TRZAX)

Photos

FIGURE 4
Overview of soft wheat (winter) (TRZAW code) on the EPPO global database website.

as is the case for the synonym *vinterhvede* included in this listing.

Then, Listing 1 provides details on the representation of the TRZAX code, which is represented as subclass of the 1TRZG code (*Triticum*). The ontological representation of this code includes almost the same properties as those described for the TRZAW code, but also details about its taxonomy (`has_taxonomy`) and taxonomy level (`has_taxonomy_level`). Moreover, this code is linked to the TRZAW code by means of the `sio:SIO_001403` (is associated with) property. It is worth mentioning that, thanks to this last link we can answer our first Competency Question. Finally, TRZAX is linked to the AGMYOR code (*Agromyza oryzae*) via the `has_pest_type_host` property, which means that TRZAX is the host of AGMYOR.

Lastly, in Listing 1, the AGMYOR code is defined as a subclass of the 1AGMYG code (*Agromyza*). The ontological representation of AGMYOR includes all the properties described for the TRZAX code. Moreover, it includes the `has_host_type_host` relationship to represent that TRZAX is the pest for which AGMYOR is relevant; that is, the inverse relationship of the property previously defined above with `has_pest_type_host`. In addition, the AGMYOR code is linked to a specific categorization status via the `has_categorization` property. This categorization is defined at the end of this listing as a subclass of the `Categorization_Status` class and is linked to the `QuarantinePest` categorization list by means of the `has_status` property. Finally, this categorization status also includes information about its categorization continent, (`has_categorization_continent`), country (`has_categorization_country`), country's iso code (`has_categorization_iso_code`), q list (`has_categorization_q_list`), and year it was added (`has_categorization_year_added`).

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix eppo: <https://ontology.basf.net/ontology/BASF/Bioscience/
      EPPO/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix oboInOwl: <http://www.geneontology.org/formats/oboInOwl#> .
@prefix sio: <http://semanticscience.org/resource/> .

eppo:TRZAW rdf:type owl:Class ;
    rdfs:subClassOf eppo:EPPONonTaxonomicCode ,
        [ rdf:type owl:Restriction ;
        owl:onProperty obo:BFO_0000050 ;
        owl:someValuesFrom eppo:3SWHC ] ;
        rdfs:label "soft wheat (winter)"@la ;
        skos:altLabel "Triticum aestivum"@la ;
        dc:identifier "TRZAW" ;
        dcterms:created "2002-03-05T00:00:00"^^xsd:dateTime ;
        dcterms:modified "2015-04-13T17:58:00"^^xsd:dateTime ;
        oboInOwl:hasDbXref obo:NCBITaxon_4565 ;
        oboInOwl:hasExactSynonym "vinterhvede"@da ,
            "winter wheat"@en ,
            "wintertarve"@nl ;
        eppo:has_eppo_type eppo:NTX ;
        eppo:is_active "true"^^xsd:boolean .

[ rdf:type owl:Axiom ;
   owl:annotatedSource eppo:TRZAW ;
   owl:annotatedProperty oboInOwl:hasExactSynonym ;
   owl:annotatedTarget "vinterhvede"@da ;
   dcterms:created "2017-07-11T22:41:00"^^xsd:dateTime ;
   eppo:is_active "true"^^xsd:boolean ] .

eppo:TRZAX rdf:type owl:Class ;
    rdfs:subClassOf eppo:1TRZG ,
    [ rdf:type owl:Restriction ;
        owl:onProperty sio:SIO_001403 ;
        owl:someValuesFrom eppo:TRZAW ] ,
    [ rdf:type owl:Restriction ;
```

```
        owl:onProperty eppo:has_pest_type_host ;
        owl:someValuesFrom eppo:AGMYOR ] ;
    dc:identifier "TRZAX" ;
    rdfs:label "Triticum aestivum"@la ;
    skos:altLabel "Triticum sativum"@la ,
           "Triticum vulgare"@la ;
    dc:identifier "TRZAX" ;
    dcterms:created "2002-02-03T00:00:00"^^xsd:dateTime ;
    dcterms:modified "2002-02-03T00:00:00"^^xsd:dateTime ;
    oboInOwl:hasDbXref obo:NCBITaxon_4565 ;
    oboInOwl:hasExactSynonym "Saatweizen"@de ,
        "bread wheat"@en ;
    eppo:has_eppo_type eppo:PFL ;
    eppo:has_taxonomy eppo:Species ;
    eppo:has_taxonomy_level 9 ;
    eppo:is_active "true"^^xsd:boolean .

eppo:NTX rdf:type owl:Class ;
    rdfs:subClassOf eppo:EPPOCodeType ;
    dc:identifier "NTX"@en ;
    rdfs:label "Non taxonomic"@en .

eppo:PFL rdf:type owl:Class ;
    rdfs:subClassOf eppo:EPPOCodeType ;
    dc:identifier "PFL"@en ;
    rdfs:label "Plant"@en .

eppo:Species rdf:type owl:Class ;
    rdfs:subClassOf eppo:EPPOCodeTaxonomyLevel ;
    oboInOwl:hasDbXref obo:NCBITaxon_species ;
    rdfs:label "Species"@en .

EPPO:AGMYOR rdf:type owl:Class ;
    rdfs:subClassOf eppo:1AGMYG ,
    [ rdf:type owl:Restriction ;
        owl:onProperty eppo:has_categorization ;
        owl:someValuesFrom eppo:
      Categorization_AGMYOR_QuarantinePest_US ] ,
    [ rdf:type owl:Restriction ;
        owl:onProperty eppo:has_host_type_host ;
        owl:someValuesFrom eppo:TRZAX  ] ;
    dc:identifier "AGMYOR" ;
    dcterms:created "2002-11-05T00:00:00"^^xsd:dateTime ;
    dcterms:modified "2002-11-05T00:00:00"^^xsd:dateTime ;
    oboInOwl:hasExactSynonym "Japanese rice leaf miner"@en ,
        "agromyze du riz"@fr ;
    rdfs:label "Agromyza oryzae"@la ;
    skos:altLabel "Agromyza oryzella"@la ,
        "Oscinis oryzae"@la ,
        "Oscinis oryzella"@la ;
    eppo:has_eppo_type EPPO:GAI ;
    eppo:has_taxonomy EPPO:Species ;
    eppo:has_taxonomy_level 8 ;
    eppo:is_active "true"^^xsd:boolean .

eppo:Categorization_AGMYOR_QuarantinePest_US rdf:type owl:Class ;
    rdfs:subClassOf EPPO:Categorization_Status ,
    [ rdf:type owl:Restriction ;
        owl:onProperty eppo:has_status ;
        owl:someValuesFrom EPPO:QuarantinePest ] ;
    eppo:has_categorization_continent "America" ;
    eppo:has_categorization_country "United States of America";
    eppo:has_categorization_iso_code "US";
    eppo:has_categorization_q_list "X" ;
    eppo:has_categorization_year_added 1994 .
```

Listing 1 Excerpt from the EPPO ontology representing the TRZAW code.

5. Adoption of the ontology

The EPPO ontology is a first step to align the whole Agricultural Solutions division on a similar vocabulary need. In BASF, we have four main agricultural focus areas: Crop Protection, Seed and Traits, Vegetable Seeds, and Digital Farming. By means of the EPPO ontology, we align these departments to work on a common vocabulary when referring to organisms.

Currently, the EPPO ontology is being used as a key element of different applications, including Bioregister. Dotmatics' Bioregister is a Web-based application for registering sequence-based, chemically modified and structure-less biological materials, allowing biologics discovery organizations to ensure entity uniqueness and protect their intellectual property. Bioregister supports management of a broad set of biological materials, including DNA, RNA, peptides and proteins, antibodies, conjugates, non-natural peptides and nucleotides, plasmids, cell lines, and user-defined entities. It also enables users to record batches and samples for these entities, purification and expression information, and other protein production data.

When users enter, for example, a new microorganism record in the application, it needs to be associated to a plant or pest. In other previous applications, the reference to these terms was manually added using a free text input area, so different terms were being used to refer to the same concept. Even when it was agreed that the EPPO codes should be used instead, there were still plenty of errors as users could inadvertently misspell the codes or use different names to refer to the same concept. Having such naming and format heterogeneity, as well as mistaken data, led to inefficiencies when exploiting Bioregister data for further analysis purposes.

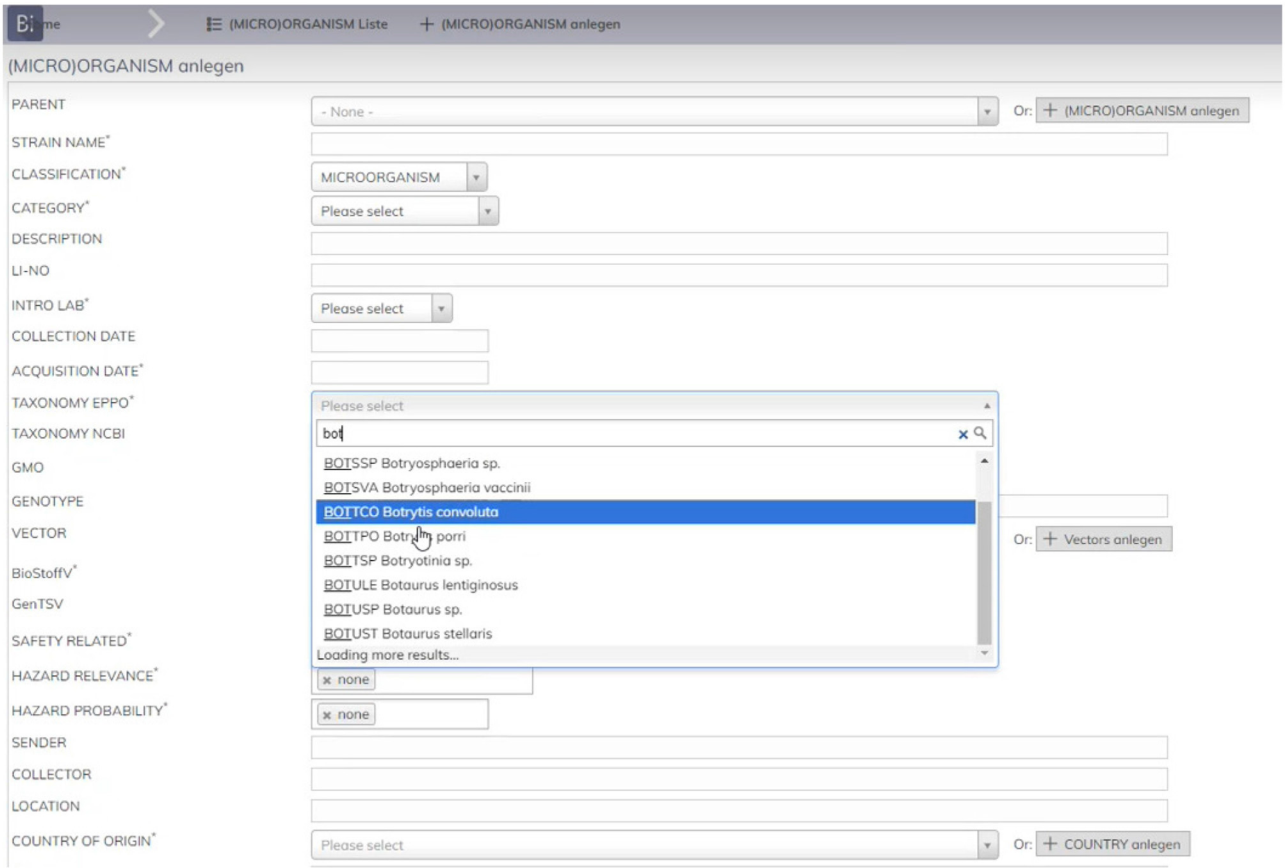
To prevent users from making errors when inserting EPPO codes, the latest version of Bioregister uses the ontology. As seen in Figure 5, Bioregister's interface has a dropdown list for users to select a specific term from the EPPO ontology. To populate such dropdown list, the application consumes the EPPO ontology through a specific API call, so that the latest version is always available, and the terms that appear in the list are dynamically updated based on what users have entered in the text area. It is worth mentioning that, to facilitate the consumption of the EPPO ontology, we have configured a REST API service which offers a whole set of generic API calls which can be used by other applications. In addition, it is worth remembering that microorganisms are just one entity example that can be included in Bioregister. Therefore, EPPO codes for the associated organism are also used for other entities such as plants or the donor organisms for constructs, enzymes, cell lines, among others.

Since the knowledge represented by the EPPO ontology is pertinent to different types of users, with different background and different IT skills, the consumption via APIs may not be enough to ensure the access to the information. Therefore, another way users consume the EPPO ontology is by means of our internal OLS. This way users may search and navigate across the different concepts when looking for information relevant to their work.

Finally, we are reusing the ontology in the development and enrichment of internal ontologies, such as for example the BASF Crop Protection Experiments ontology. This ontology aims to represent the process that is carried out in our labs to design, plan, prepare, execute, and assess experiments to identify new active ingredients or traits protecting crops against pests and diseases.

6. Conclusions and future work

In this work, we presented the ontology we developed to represent the EPPO coding system. The ontology includes the



(MICRO)ORGANISM anlegen

PARENT	- None -	Or: + (MICRO)ORGANISM anlegen
STRAIN NAME*		
CLASSIFICATION*	MICROORGANISM	
CATEGORY*	Please select	
DESCRIPTION		
LI-NO		
INTRO LAB*	Please select	
COLLECTION DATE		
ACQUISITION DATE*		
TAXONOMY EPPO*	Please select bot BOTSSP Botryosphaeria sp. BOTSVA Botryosphaeria vaccinii BOTTCO Botrytis convoluta BOTTPO Botrytis porri BOTTSP Botryotinia sp. BOTULE Botaurus lentiginosus BOTUSP Botaurus sp. BOTUST Botaurus stellaris Loading more results...	
TAXONOMY NCBI		
GMO		
GENOTYPE		
VECTOR		Or: + Vectors anlegen
BioStoffV*		
GenTSV		
SAFETY RELATED*		
HAZARD RELEVANCE*	× none	
HAZARD PROBABILITY*	× none	
SENDER		
COLLECTOR		
LOCATION		
COUNTRY OF ORIGIN*	Please select	Or: + COUNTRY anlegen

FIGURE 5
Adoption of the EPPO ontology in Bioregister.

data available in several files from the EPPO Global Database and also the information provided in its REST API. In addition, we defined a granular hierarchy of the EPPO Code phytosanitary categorizations that represents the general categories defined in the EPPO lists, European Union lists, and beyond. Finally, we enriched the ontology with NCBITaxon cross-references to allow consuming further information from such knowledge base.

During the development of this work, we have learned several lessons that will help us to improve our ontology developments in the future. First, although the automatic development of ontologies is a valuable method for representing huge data sources, the intervention of domain experts during the process is essential. In our experience the experts have been key to define the requirements, develop the competency questions, and validate both conceptual model and the results obtained after the execution of our Python package. Several relationships that were not implicitly defined in the EPPO coding system have been defined by our experts, and as a result we have a more granular categorization of EPPO Code phytosanitary categorizations. Second, the ontology development is a process that is time and resource intensive, but this is insignificant compared to what we save up by having only one source of EPPO codes. Third, adoption of ontology has not been an easy path in our company because, as happens in most organizations whenever a new technology appears, there is a certain skepticism about the results that can be obtained by applying it. However, more and more departments are being encouraged to use it to improve their processes.

Despite the advantages of reusing traditional upper-level ontologies (e.g., DOLCE, Masolo et al., 2002) to ease interoperability, we are not reusing them at BASF. The main reason for such a decision is that this kind of monolithic ontologies introduce strong commitments that make it difficult to represent in a lightweight manner our domains of interest. However, parallel to the development of the EPPO ontology, a new work team was formed to develop BASF core ontologies that encapsulate the terms and relationships that are of crucial relevance to the company and that will path the way to facilitate our internal interoperability. Therefore, as part of future work, we will improve the representation of categorization locations of the ontology. For this purpose, we plan to reuse our recently released BASF Core Locations ontology which represents the geographical locations across BASF including administrative areas (such as countries, cities, among others) and location of points of interest (such as production plants and sites, among others). Therefore, we can reuse the concepts from that core ontology to represent countries, regions and ISO country codes instead of representing them as string values as currently done in the classes defined as part of the EPPO Categorization Status.

By having such concepts linked to our ontology, we will be able to get more details to, for example, infer in which cities the phytosanitary categorization is applicable and therefore know in which of our production plants we have to take special care in case of a pest. We can also take advantage of the geometric values contained in the core ontology to have a map that can provide us with alerts on the categorizations in a customized way for the points of interest relevant to our company.

Within BASF, the Biosafety function has oversight on the use of all types of biological material in facilities with the aim of protecting human health and the environment and to prevent their misuse (biosecurity) while ensuring compliance with regulatory and company requirements. Hence, a possible future direction is the development of a Risk Group Classification ontology aimed to represent not only the list of phytosanitary categorizations included in the EPPO ontology, but also data whether organisms are regulated as human or animal pathogens in selected countries around the world. Having the regulatory categorization of plant, human and animal pathogens in a single data source which can be easily queried allows to identify in a single effort the applicable government regulations pertaining to these organisms in a certain geography, instead of having to manually consult various public/external data sources, as well as supports aligned biorisk management approaches across different BASF sites and countries.

Additionally, there are plans to reuse the ontology in applications that are used internally such as Ceres (for managing the inventory of biological materials in our R&D laboratories and greenhouses) or PhenomeOne (for managing the entire plant research information of the organization, providing support for all the stages of our experimental processes). Finally, since ontologies can change, we will implement a monitoring and updating mechanism to track NCBITaxon updates. This way, if something changes in that ontology, our EPPO ontology will be aligned with it.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://gd.eppo.int.

Author contributions

AA-B, JB-D, BM, KH, and DB: conceptualization and design. AA-B, JB-D, TC, CC, and NP: development. PE-A: writing—original draft preparation. PE-A, IE-G, DB, AA-B, and JB-D: writing—review and editing. PE-A, IE-G, and AA-B: supervision. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank Alexander García Castro for the conceptualization of the EPPO ontology development.

Conflict of interest

AA-B, JB-D, PE-A, and IE-G were employed by BASF Digital Solutions. DB was employed by BASF Belgium Coordination Center CommV, Innovation Center Gentm. BM was employed by BASF Corporation. KH was employed by BASF SE Data Management and Data Governance, Global Research Services APR/HP. TC was employed by TalentBay. CC and NP were employed by Linking Data SAS.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1131667/full#supplementary-material

References

Alomar, O., Batlle, A., Brunetti, J., García, R., Gil, R., Granollers, A., et al. (2015). Development and testing of the media monitoring tool med is ys for early identification and reporting of existing and emerging plant health threats. *EPPO Bull.* 45, 288–293. doi: 10.1111/epp.12209

Alomar, O., Batlle, A., Brunetti, J. M., García, R., Gil, R., Granollers, T., et al. (2016). Development and testing of the media monitoring tool MedISys for the monitoring, early identification and reporting of existing and emerging plant health threats. *EFSA Support. Publ.* 13, 1118E. doi: 10.2903/sp.efsa.2016.EN-1118

Bastian, F., Overton, J., Dietze, H., Mungall, C., Midford, P., Duncan, B., et al. (2013). *NCBITaxon Ontology*. doi: 10.5281/zenodo.7676251

Chávez-Feria, S., García-Castro, R., and Poveda-Villalón, M. (2022). "ChowLK: from UML-based ontology conceptualizations to OWL," in *European Semantic Web Conference* (Springer), 338–352.

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., et al. (2013). The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.* 54, e1. doi: 10.1093/pcp/pcs163

Côté, R. G., Jones, P., Apweiler, R., and Hermjakob, H. (2006). The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. *BMC Bioinformatics* 7, 97. doi: 10.1186/1471-2105-7-97

Damos, P., Karampatakis, S., and Bratsas, C. (2017). Representing and integrating agro plant-protection data into semantic web through a crop-pest ontology: the case of the Greek Ministry of Rural Development and Food (GMRDF) Ontology. *IOBCWPRS Bull.* 123, 122–127.

Darnala, B., Amardeilh, F., Roussey, C., and Jonquet, C. (2021). "Crop Planning and Production Process Ontology (C3PO), a new model to assist diversified crop production," in *Integrated Food Ontology Workshop (IFOW'21) at the 12th International Conference on Biomedical Ontologies (ICBO).*

European Food Safety Authority, Mannino, M. R., Larenaudie, M., Patrick Linge, J., Candresse, T., Jaques Miret, J. A., et al. (2021). *Horizon Scanning for Plant Health: Report on 2017-2020 Activities*. Technical report, Wiley Online Library.

European and Mediterranean Plant Protection Organization (2022). PP 1/248 (3) Harmonized system for classification and coding of the uses of plant protection products. *EPPO Bull.* 52, 17–24. doi: 10.1111/epp.12789

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic Acids Res.* 40, D136–D143. doi: 10.1093/nar/gkr1178

Fernandez-Lopez, M., Poveda-Villalon, M., Suarez-Figueroa, M. C., and Gomez-Perez, A. (2019). Why are ontologies not reused across the same domain? *J. Web Semant.* 57, 100492. doi: 10.1016/j.websem.2018.12.010

Grüninger, M., and Fox, M. S. (1995). "Methodology for the design and evaluation of ontologies," in *Workshop on Basic Ontological Issues in Knowledge Sharing* (Montreal).

Iglesias-Molina, A., Bernabe-Diaz, J. A., Deshmukh, P., Espinoza-Arias, P., Fernandez-Izquierdo, A., Ponce-Bernabe, J. M., et al. (2022). *Ontology Management in an Industrial Environment: The BASF Governance Operational Model for Ontologies (GOMO)*. Zenodo.

IPPC Secretariat, Gullino, M., Albajes, R., Al-Jboory, I., Angelotti, F., Chakraborty, S., et al. (2021). *Scientific Review of the Impact of Climate Change on Plant Pests–A Global Challenge to Prevent and Mitigate Plant Pest Risks in Agriculture, Forestry and Ecosystems*. FAO on behalf of the IPPC Secretariat.

Jackson, R. C., Balhoff, J. P., Douglass, E., Harris, N. L., Mungall, C. J., and Overton, J. A. (2019). ROBOT: a tool for automating ontology workflows. *BMC Bioinformatics* 20, 407. doi: 10.1186/s12859-019-3002-3

Lacasta, J., Lopez-Pellicer, F. J., Espejo-García, B., Nogueras-Iso, J., and Zarazaga-Soria, F. J. (2018). Agricultural recommendation system for crop protection. *Comput. Electron. Agric.* 152, 82–89. doi: 10.1016/j.compag.2018.06.049

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., and Schneider, L. (2002). The wonder web library of foundational ontologies. *WonderWeb Deliverable D* 17, 2002.

Noy, N. F., and Musen, M. A. (2003). Ontology versioning as an element of an ontology-management framework. *IEEE Intell. Syst.* 19, 6–13. doi: 10.1109/MIS.2004.33

Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., and García-Castro, R. (2022). LOT: an industrial oriented ontology engineering framework. *Eng. Appl. Artif. Intell.* 111, 104755. doi: 10.1016/j.engappai.2022.104755

Roy, A.-S. (2019). *EPPO Codes - An Overview*. Technical report, European and Mediterranean Plant Protection Organization.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data Knowledge Eng.* 25, 161–197.

Tartir, S., Arpinar, I. B., and Sheth, A. P. (2010). "Ontological evaluation and validation," in *Theory and Applications of Ontology: Computer Applications*, 115–130.

Tudorache, T., Nyulas, C., Noy, N. F., and Musen, M. A. (2013). WebProtégé: a collaborative ontology editor and knowledge acquisition tool for the web. *Semant. Web* 4, 89–99. doi: 10.3233/SW-2012-0057

frontiers | Frontiers in Artificial Intelligence

TYPE Conceptual Analysis
PUBLISHED 04 July 2023
DOI 10.3389/frai.2023.1137961

Ontological how and why: action and objective of planned processes in the food domain

Damion Dooley[1]*† and Tarini Naravane[2]†

[1]Centre for Infectious Disease Genomics and One Health, Simon Fraser University, Burnaby, BC, Canada,
[2]Biological Systems Engineering, University of California, Davis, Davis, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Catherine Roussey,
INRAE Occitanie Montpellier, France

REVIEWED BY
Patrice Buche,
Institut National de recherche pour l'agriculture, l'alimentation et l'environnement (INRAE), France
Sylvie Despres,
Université Sorbonne Paris Nord, France
María Poveda-Villalón,
Universidad Politécnica de Madrid, Spain

*CORRESPONDENCE
Damion Dooley
✉ damion_dooley@sfu.ca

†These authors have contributed equally to this work

RECEIVED 05 January 2023
ACCEPTED 05 June 2023
PUBLISHED 04 July 2023

CITATION
Dooley D and Naravane T (2023) Ontological how and why: action and objective of planned processes in the food domain.
Front. Artif. Intell. 6:1137961.
doi: 10.3389/frai.2023.1137961

The computational modeling of food processing, aimed at various applications including industrial automation, robotics, food safety, preservation, energy conservation, and recipe nutrition estimation, has been ongoing for decades within food science research labs, industry, and regulatory agencies. The datasets from this prior work have the potential to advance the field of data-driven modeling if they can be harmonized, but this requires a standardized language as a starting point. Our primary goal is to explore two interdependent aspects of this language: the granularity of process modeling sub-parts and parameter details and the substitution of compatible inputs and processes. A delicate semantic distinction—categorizing planned processes based on the objectives they seek to fulfill vs. categorizing them by the actions or mechanisms they utilize—helps organize and facilitate this endeavor. To bring an ontological lens to process modeling, we employ the Open Biological and Biomedical Ontology Foundry ontological framework to organize two main classes of the FoodOn upper-level material processing hierarchy according to objective and mechanism, respectively. We include examples of material processing by mechanism, ranging from abstract ones such as "application of energy" down to specific classes such as "heating by microwave." Similarly, material processing by objective—often a transformation to bring about materials with certain qualities or composition—can, for example, range from "material processing by heating threshold" to "steaming rice".

KEYWORDS

food processing, ontology, mechanism, action, data specification, material processing

1. Introduction

The post-harvest treatment of food up to the point of consumption from both industrial and domestic food preparation perspectives is an active area of research that is not yet comprehensively covered by an integrated set of ontologies. Here, we propose for discussion, as part of a larger life-sciences family of ontologies, the basic terms required for a standardized process ontology that can enable and integrate data-driven analysis of research datasets on the one hand and (with data at the relevant resolution and data size) support dynamic process control applications on the other hand. Formalized language is required to integrate what have often been siloed food composition datasets (FCD) containing foods that result from simple processes such as boiling, freezing, and roasting. Standardized language is a prerequisite to the manual or automated alignment of different food entities across FCD datasets. For example, the nutritional content of frozen carrots can only be compared across datasets if the experimental protocols for storing, soaking, blanching, boiling, and (flash) freezing processes are comparable (Hinojosa-Nogueira et al., 2021; Westenbrink et al., 2021).

Ontologies can be used to normalize the comparable portions of data selected from the body of scientific literature on food processing so that data-driven analysis and models can be used to address a range of research questions/hypotheses on the causal factor(s) driving sensory and nutritive effects.

This process ontology work also aims to support dynamic process control by providing a framework for describing process input and output phenotype objective thresholds that can trigger mechanism start/stop/pause operations and by providing a framework for choosing among comparable mechanisms to achieve an objective. The ontology has been designed to differentiate between the objectives and mechanisms of a process and to address food processing at both macro (food product) and molecular scale transformations. The latter is especially challenging to describe in food science literature datasets (a task similar to material science engineering modeling). We provide a macro/micro transformation example model that shows the parallels between the food entity and molecular resolutions. This framework addresses the various details of a process to support various decision-making methods for dynamic process control, ranging from simple inferences/linear relationships to more complex ML models.

Given the natural context of food as mainly derived from organisms, our process ontology leverages the framework established by the Open Biological and Biomedical Ontology Foundry (OBO) consortium of ontologies (Jackson et al., 2021), which focuses on life science research. In this study, the modeling of natural physio-chemical or biologically rooted processes can be found in places such as the Gene Ontology's cellular metabolic process [GO:0008152] branch classes (including, for example, fermentation, an enzyme-catalyzed process), which are triggered when some combination of materials and/or environmental context aligns. Unplanned processes can be controlled by planned processes that exhibit human or computer agency/intentionality. To organize processes that satisfy various objectives in transforming things, OBO's Ontology for Biomedical Investigations (OBI) (Bandrowski et al., 2016) introduced the "planned process" class [OBI:0000011], which contains processes that execute a "plan specification" and include a set of instructions and/or objectives.

A recent paper (Dooley et al., 2022) covers a gap analysis of the technical side of modeling processes using W3C OWL ontologies (SOSA, SSN, PO2, and OWL-Time) in comparison to an OBO

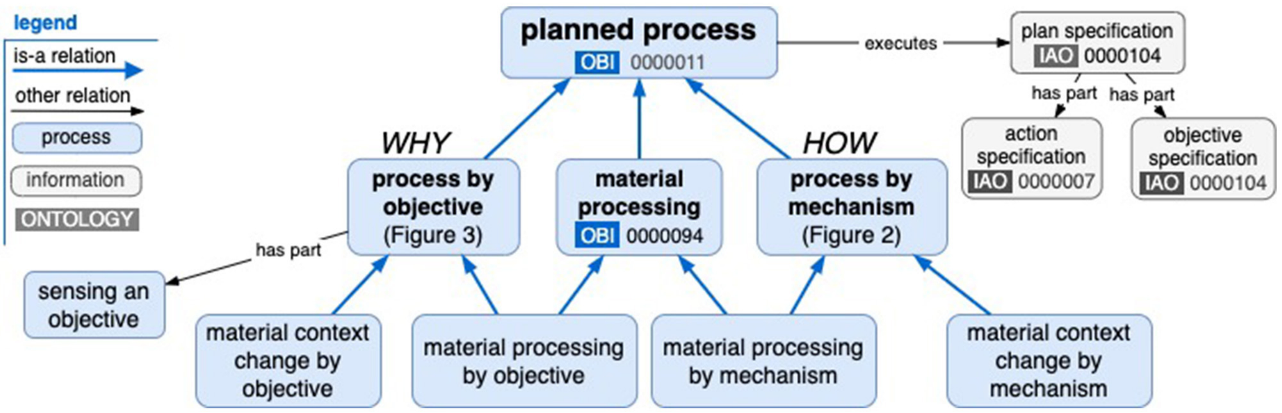


FIGURE 1
A new material processing hierarchy organized under a planned process, with new "process by objective" and "process by mechanism" branches.

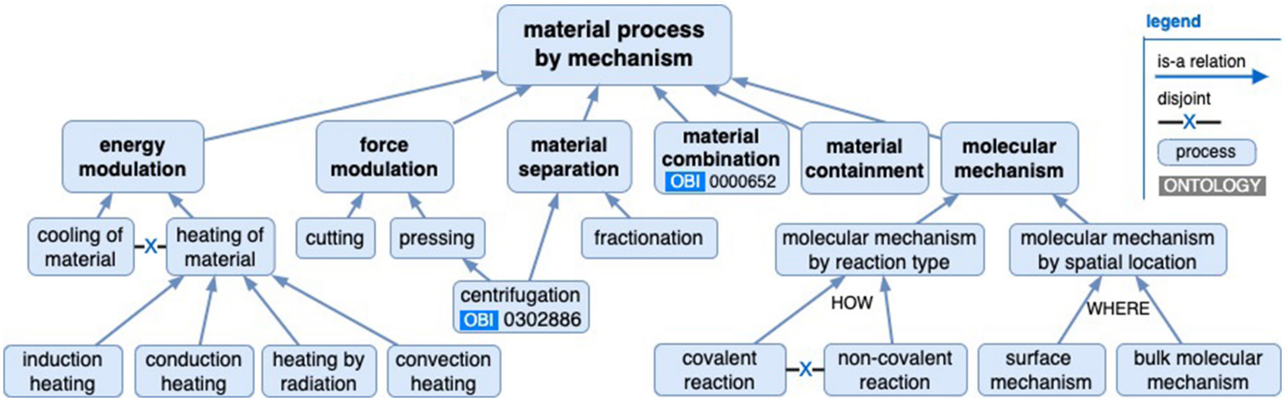


FIGURE 2
A material processing by mechanism hierarchy with example subclasses, in which no absolute end-point material qualities are specified.

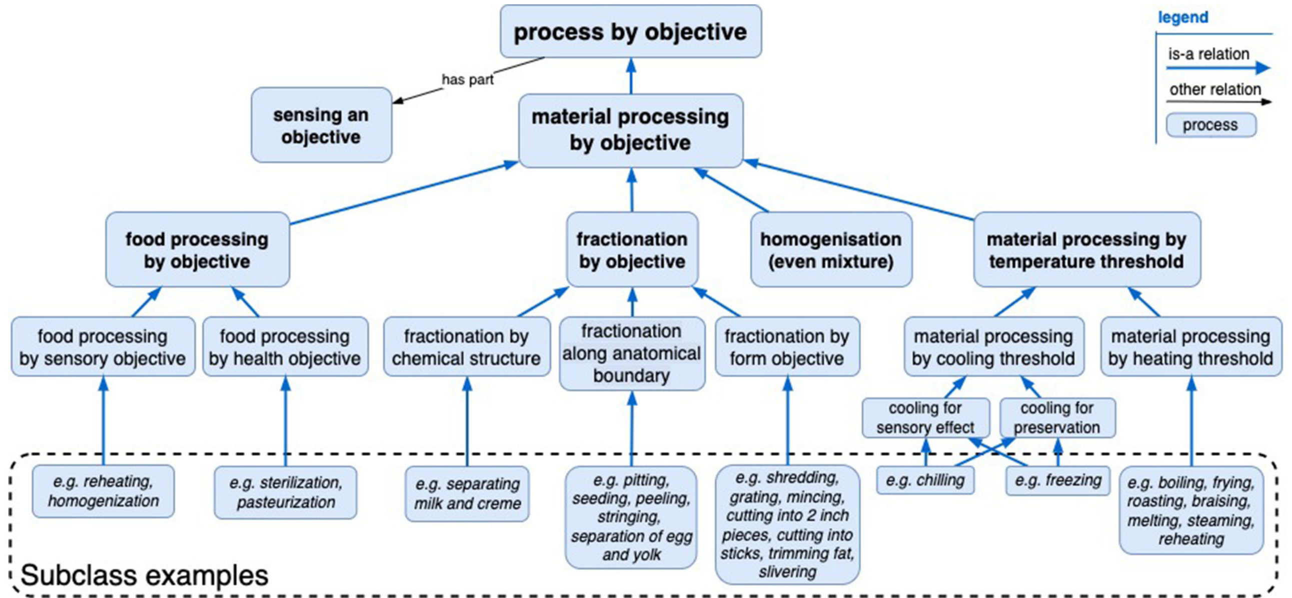

process by objective
has part
sensing an objective
material processing by objective
legend
is-a relation
other relation
process
food processing by objective
fractionation by objective
homogenisation (even mixture)
material processing by temperature threshold
food processing by sensory objective
food processing by health objective
fractionation by chemical structure
fractionation along anatomical boundary
fractionation by form objective
material processing by cooling threshold
material processing by heating threshold
cooling for sensory effect
cooling for preservation
e.g. reheating, homogenization
e.g. sterilization, pasteurization
e.g. separating milk and creme
e.g. pitting, seeding, peeling, stringing, separation of egg and yolk
e.g. shredding, grating, mincing, cutting into 2 inch pieces, cutting into sticks, trimming fat, slivering
e.g. chilling
e.g. freezing
e.g. boiling, frying, roasting, braising, melting, steaming, reheating
Subclass examples

FIGURE 3
A hierarchy of "material processing by objective" classes and some lower-level examples. Each has a plan specification containing an objective specification that supplies a completion metric.

Foundry ontology approach and recommendations, implemented mainly in OBO Foundry's FoodOn food ontology (Dooley et al., 2018) for extending OBO with some select relationships and classes adapted from the aforementioned ontologies to fill the gap. The paper details the OBI "planned process" related classes and relations and discusses how experimental independent and dependent variables, observations, and characteristics of materials could be structured in a multi-step process model. A brief discussion of measurement data properties is included, but in that (and current) work, we avoid focusing on this topic and note that upcoming OBO work will recommend knowledge graph data structures for measurement values. That paper finishes with a simple recipe model that illustrates ingredient input and output relations at work in a sample selection of food processes but skirts the issue of organizing a hierarchy of food processes; our new work focuses on this topic.

A plan specification can have one or more "action specifications" [IAO:0000007] parts that directly or indirectly control the input material's environmental parameters, such as container pressure, kinetic or thermal energy exposure, or the addition of chemicals or biological substances. An action specification might be to operate a tool or device setting or control to some effect or to give hands-on instruction to an operator to shape a material directly or combine materials. It may reference other planned processes or directly control (via duration, catalysts, or energy supply) an unplanned physio-chemical process. Natural fermentation would be considered unplanned, but a planned process can harness it through action specification(s), devices, and subprocess stages. Other examples are the application of force to a material or to a blade in the material; introducing bacteria to a food substance; controlling atmospheric storage conditions for food; or allowing the fruit to ripen before harvest or consumption (Osorio et al., 2013).

In time, the effect of environmental interventions (whether constant or in flux) yields physical or chemical changes in material input(s) that satisfy process objectives. A material processing "objective specification" is often an expression of the quality(ies) or phenotype(s) of the output material, such as "water at 100 degrees Celsius", which is the causal result of the process. Other examples are sensory, logistical, food safety, or food formulation functional objectives. In short, an objective is an expression of some desired state of affairs, and the "Process by Objective" class, thus, necessarily includes such an expression either as a final output specification to reach or by a formula of operating parameters. The recognition that an objective has been attained (whether by a human or a device) can be a component objective of a larger process.

2. Methodology and results

2.1. Process terms

In OBO, currently, there are no "convenience classes" for organizing processes by action or objective, so our proposal involves adding those new process terms and underlying ones within an appropriate OBO ontology. FoodOn could temporarily accommodate them, but OBO's best practice entails consulting about the possible adoption of mid/upper-level terms by the curators of OBI or the in-development Core Ontology for Biology and Biomedicine (COB) (Core Ontology for Biology and Biomedicine, n.d.), which is taking on commonly used OBO terms. Although specific processes such as boiling are examined here in

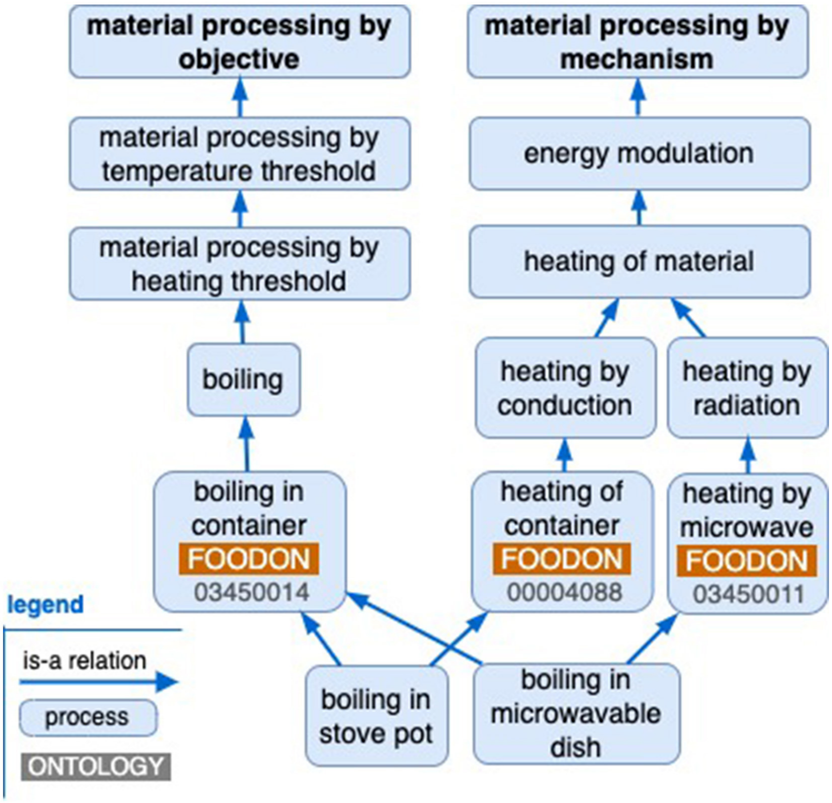


FIGURE 4
A sketch of polyhierarchical process dependencies.

the food context, they are also often applicable to other domains such as manufacturing and laboratory procedures and are best curated in a general-purpose ontology from which FoodOn can draw. Note that in this study, the "is-a" relationship in the legend refers to OWL rdfs:subclassOf. Additionally, all illustrated relations are RO or OBI ontology relations and have their domain and range constraints held in those ontologies (such as RO "has quality" range "quality" [PATO:0000001]).

FoodOn has an existing "food transformation process" branch, which will be reorganized according to the scheme proposed below. The branch is managed according to a common OBO term maintenance pattern as a spreadsheet template (FoodOn Robot Tables, 2023), which is periodically converted into a stand-alone ontology import file. Figure 1 offers an overview of the new proposed hierarchy with new "process by objective" and "process by mechanism" classes alongside the existing OBI "material processing" term. The "material context change by objective" and "material context change by mechanism" classes cover both packaging and moving of material entities (to some objective location or by some mechanism of transport), but they are not detailed here.

The "material processing by mechanism" branch outlined in Figure 2 covers the application of force or energy and combining materials (and includes some example subclasses). In this study, the relative change effected by a process will modulate a material quality, such as by reducing particle size, changing temperature, or adding a new quality, but it will not specify an absolute threshold upon which to complete the process. These processes continue unless some inherent process limit occurs, such as an exhausted resource or, with mixing miscible liquids, if maximum homogenization is reached.

The "material processing by objective" branch outlined in Figure 3 includes complete processes when one or more objectives are satisfied. This can involve objectives that are expressed as threshold qualities of a material, such as a turkey with a core temperature of 70°C. Alternatively, objectives may be expressed as characteristics of the process—for example, its duration, energy, or amount of catalyst consumed—which are a proxy for predicted material outcomes. When applied to food products, terms such as "chilling" may have highly industry-specific objective semantics, such as the chilling of animal products (Temperatures and Chilling and Freezing Procedures, 2023), which could be formalized in the ontology. The proposed material processing by objective hierarchy does not preclude objective specifications, so a reasoner should be able to infer that material processing by objective classes falls under more general process mechanism classes, for example, "material processing by cooling threshold" as a subclass of "cooling of the material," or "fractionation by objective" as a subclass of "fractionation".

Positioning of a process by objective (for example, bringing a liquid to its boiling point) and by mechanism (for example,

TABLE 1 A new process hierarchy based on objective and mechanism branches.

Label	Definition	Notes
Planned process (OBI)	A process that realizes a plan that is the concretization of a plan specification.	This term is from OBI. Paraphrasing: a process that executes a plan specification.
Process by mechanism	A planned process that has one or more action specification parts in its plan specification that control a mechanism.	A convenience class for organizing processes by their physical mechanism or digital algorithm. An action may be physical, such as pushing a button or setting a dial, or it may be about running some software.
Material processing (OBI)	A planned process that results in physical changes in a specified input material.	More than one input material may be involved. Note that ENVO's similar "material transformation process" is unplanned.
Material processing by mechanism	Material processing has one or more action specifications in its plan specification.	This should also be inferred under "process by mechanism." Here, action specifications directly or indirectly control a material input's environmental parameters.
Energy modulation	A material processing mechanism in which energy is removed from or added to a material entity.	
Heating of material	An energy modulation in which thermal energy (heat) is applied to a material or its environment.	
Cooling of material	An energy modulation in which thermal energy (heat) is removed from a material or its environment.	
Force modulation	A material processing mechanism in which force is applied to a material or its environment.	
Separating material	A material processing mechanism in which materials are separated.	
Combining material	A material processing mechanism in which materials are combined.	
Molecular mechanism	Material processing is described for specific molecules in the material.	
The molecular mechanism by reaction type	A molecular mechanism is categorized by its reaction type.	
Covalent reaction	A molecular mechanism by reaction type involving a covalent reaction.	
Non-covalent reaction	A molecular mechanism by reaction type involving a non-covalent reaction.	
The molecular mechanism by spatial location	A molecular mechanism is categorized by the region in which reactions occur.	
Surface mechanism	A molecular mechanism where reactions occur at some surface boundary.	
Bulk molecular mechanism	A molecular mechanism where reactions occur throughout a mixture.	
Process by objective	A planned process that has one or more objective specification parts in its plan specification.	A convenience class under which various processes can be grouped or inferred by their objectives.
Material processing by objective	Material processing that has one or more objective specification parts in its plan specification.	This will also be inferred under "process by objective." Here, processes having equivalent objectives can be swapped.
Direct heating of container (FoodOn)	A heating container process in which the container conducts heat by being near an open flame, a hot surface, or an oven.	
Boiling	A material processing by the heating threshold in which the objective is to keep a liquid at its boiling temperature under atmospheric conditions.	
Material context change	A planned process in which the relation of the input material entity and its proximate environment changes.	
Material context change by objective	A material context change in which the objective is to change the contextual relation of the input material entity and its environment.	For example, the objective of a wrapped food or moving some food somewhere specific.
Material location change process	A material context change in which the objective is to move the input material to another location.	The ultimate location may be dynamically ascertained based on other inputs/decision points, for example, in a sorting process [this is also an Industrial Ontologies Foundry term (Kulvatunyou et al., 2022)].
Material context change by a mechanism	A material context change is when an action that changes the contextual relations of the input material entity is applied.	For example, pushing against an object may cause it to move.

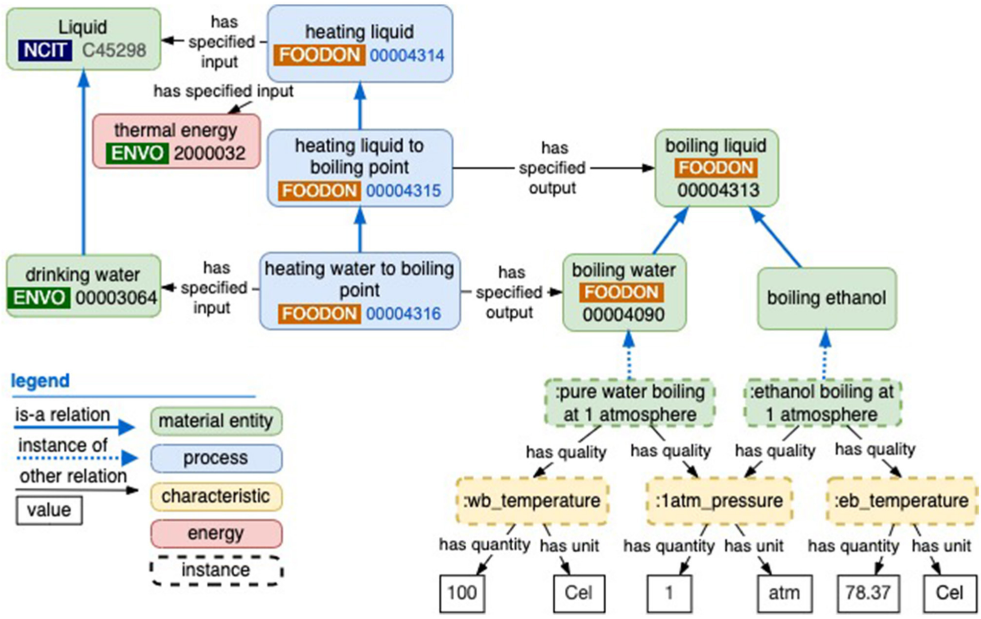


FIGURE 5
How an objective specification like "boiling water" can be expressed as a universal (class-level) output of a planned process. Boiling water is a class of boiling liquid with certain temperature and pressure characteristics, similar to other liquids like boiling ethanol.

"heating by microwave" or "direct heating of container") is shown in Figure 4, as is the example of the polyhierarchy of stove top and microwave boiling processes.

While some terms mentioned in the above figures (shown with identifiers) come from existing OBO Foundry and other ontologies, the bulk of this upper-level hierarchy must be created. New key terms are listed in Table 1, along with their definitions and notes. Our motivation for presenting these terms here is to encourage feedback in the spirit of an open-source community so that their labels and definitions can be finalized. Discussion can be held at the GitHub FoodOn issue page https://github.com/FoodOntology/foodon/issues/262 or by contacting the authors directly.

The boiling water process example illustrates the distinction between processes which have more open-ended mechanisms, and those with completion objectives. A "heating liquid" class does not include any objective, but its "heating liquid to boiling point" subclass does require a boiling liquid output. More specifically, an objective to bring some potable or "drinking water" (usually at ambient temperature) to a boiling point may require some context for that boiling, e.g., the proxy objective of it being 100 degrees Celsius (°C) at 1 atmosphere (atm) unit. Additionally, a mechanism invoked to boil this water will require a liquid container, a vector of energy, and either a "heating of the container" process or a "heating by microwave" process. As shown in Figure 5, various liquids have different boiling point temperature x atmospheric pressure objectives. There is the potential for establishing a digital library of such instants—much like the SI library of real-world entities such as the meter and the kilogram—that can be reused to express process objectives. To model the process of "heating liquid to the boiling point," one can reuse a URI that points to a "reference" instance of "boiling water" with its standard measurements of water and atmospheric conditions and a separate input instance of water with qualities that approach the standard over time (the multicomponent nature of these reference measures precludes a solution at the class level involving owl:hasValue; instead, the "has quantity" and "has unit" properties are in line with OBO's upcoming data model).

2.2. Molecular branch

The concepts of mechanism and objective apply to food at both macro and molecular levels, which gives rise to a correspondence between mechanism and objective activity at both levels. We describe the considerations for building the "molecular mechanism" branch and provide an example. This branch describes mechanisms specific to certain molecules that can be key to the molecular composition of a food's processed versions. There are at least two prominent characteristics of molecular-scale mechanisms: the chemistry of interactions between molecules and the spatial location of interactions within food material(s). Molecular interactions are either covalent (e.g., Maillard reaction) or non-covalent (including van der Waals forces, electrostatic forces, and hydrogen bonding) (Yamada, 2014) and may either occur throughout the material or be localized (Doi, 2013). Figure 8 provides a rice cooking example that identifies and differentiates various molecular mechanisms and sensor measurement concepts. Rice cooking is dominated by the molecular mechanism of starch interacting with water through different time and temperature conditions (starch comprises up to 90% of a rice kernel).

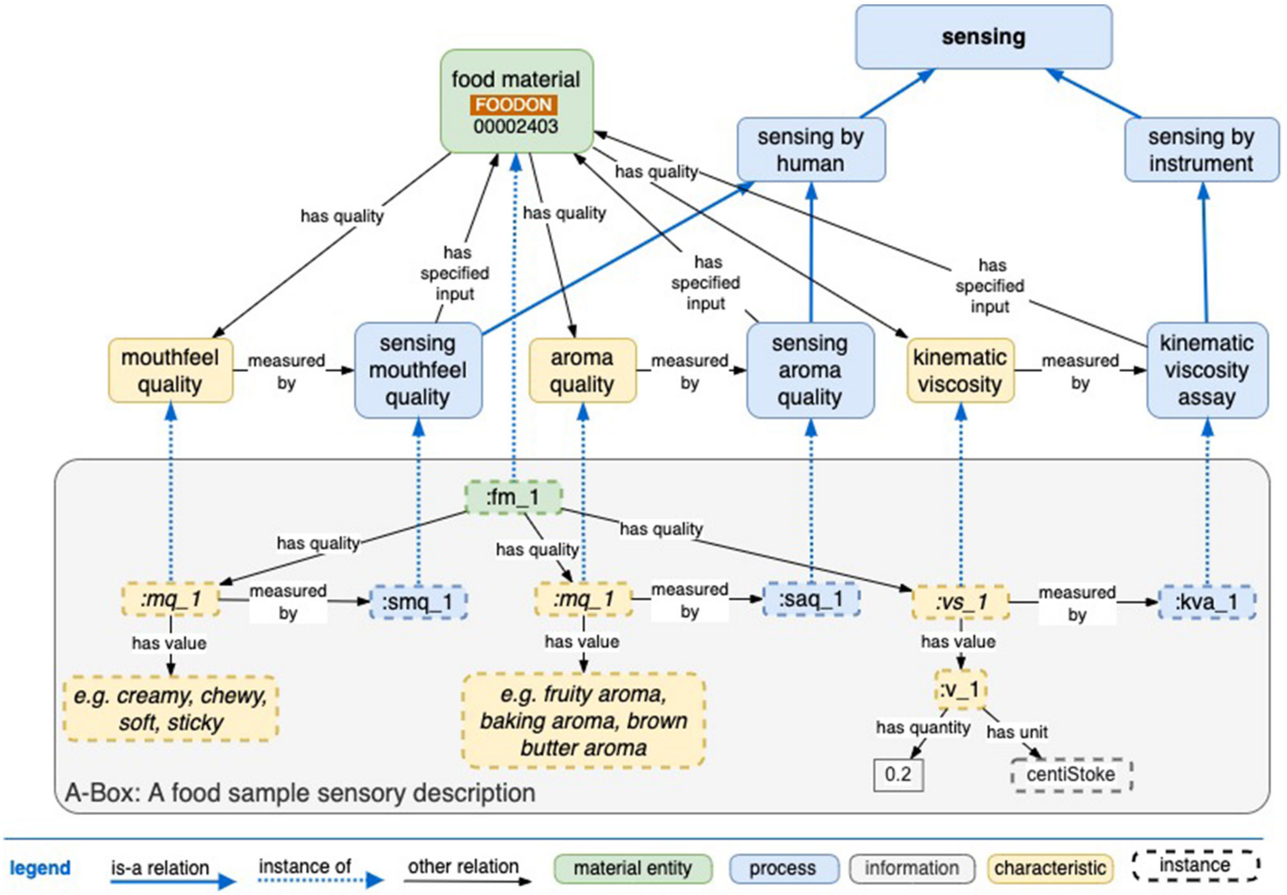


FIGURE 6
Generic schema of some sensory processes that help determine whether the objectives have been achieved during material processing.

The specific molecular processes are swelling, gelatinization, pasting, and retrogradation, all due to the hydrogen bonding interactions that occur in bulk when rice is cooked by soaking and boiling in water and then cooling. Initially, the components of starch, amylose (AM), and amylopectin (AP) are in the native "granular" state of alternating bands of amorphous and crystalline regions enabled by intramolecular hydrogen bonds. Soaking rice in water at ambient temperature sets off the gradual seepage of water into the structure at a rate proportional to the temperature. Heating this mixture of "soaked" rice and water increases the swelling of native starch granules. During this swelling process, the water creates hydrogen bonds with amylopectin and gradually disrupts the crystallinity of the granule irreversibly. This leads to the breaking of the native structure. Amylose and amylopectin leach into the water, and the gelatinization is complete. This is followed by pasting until the rice is "cooked." AP and AM reassociate as the rice cools, a process termed retrogradation (Kadam et al., 2015). This is observed as the drying-out of the rice when refrigerated or the thickening of rice porridge (congee). These mechanisms can be sensed either by humans or an instrument and are associated with several objectives, as shown in Figure 6. Specific to the example explored above, instrumental sensors indicate rheological and physical properties, while humans sense the mouthfeel qualities described as stickiness, chewiness, creaminess, etc.

2.3. Applications

The language and hierarchy of terms developed here apply to both scientific experiments and home cooking contexts, as shown in the context of the rice cooking example (Figure 7). From a domestic consumer-end recipe perspective, rice cooking often has a more formulaic approach, specifying a device and ingredient quantity, and completion is assumed by either the cooking time or by a sensory perception of mouthfeel (Naravane and Lange, 2018). From a food science perspective, rice cooking is described by the molecular mechanisms of swelling and gelatinization that specific instruments and protocols can measure. In addition, the language also addresses both macro-level and molecular-scale mechanisms, with the aim that changes in food composition can be explained at the molecular level.

The vast body of research literature on food processing addresses diverse questions to discover the sensory and nutritional profiles of processed foods due to processing conditions. Several

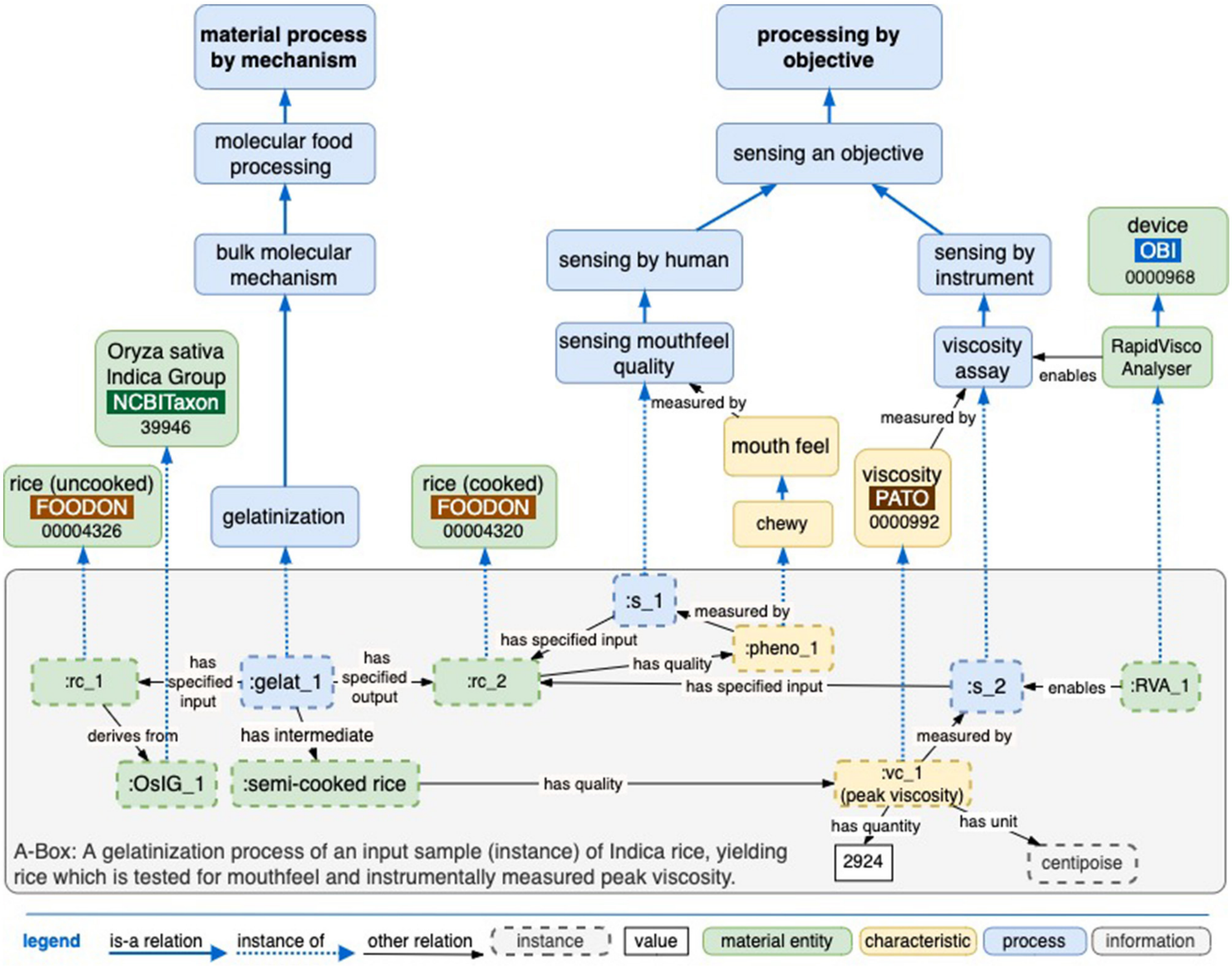


FIGURE 7
A basic model of rice cooking where completion is judged by some characteristic that is sensed by either humans or instruments. The A-Box (assertion box) expression uses the above T-Box (terminology box) ontology terms to express instances of experimental data.

experimental studies also aim to correlate objective measures with more subjective human sensory scales (Tao et al., 2020). Every experimental dataset typically explores only a few variables for specific outcomes but is taken together. These studies contribute to a vast body of research on food composition and transformative mechanisms. Integrating this experimental body requires a standardized language like this, and such large datasets have the potential for knowledge modeling, as evidenced by the models developed on traditional nutrition-focused datasets (Naravane and Tagkopoulos, 2023).

Figure 8 illustrates the rice cooking use case of applying ontology to experimental studies. Rice cooking predominantly involves the interaction of water with rice through the energy supplied in the form of heat up to the boiling point of 100°C. A progression of rice states is shown in the material entity tier of the figure. Intermediate and final process products can be measured for experimental or process control variables. This abbreviated protocol omits some steps and controls one might have, such as washing rice, using a certain cooking device, setting the cooking temperature, etc. Specifically, the "gelatinization" process has been detailed since it is essential to cooking by virtue of the water penetrating the rice's native starch crystal structure, which advances to some extent in "soaking rice", and the subsequent breakdown of crystal structure requires "heating of rice in water". However, it will take more detailed modeling to address dried rice types such as having a pertinent kind of starch crystal formation, having husks removed, and water temperature factors to replace this simplified protocol.

Various material entities are observable in both "domestic" cooking and scientific experiments, while instrumental measures (such as peak viscosity and glass transition) that capture certain molecular states are specific to a scientific context. An example of dynamic process control involves modifying time and temperature conditions to affect two outcomes: the recrystallization of rice (which is associated with glycemic index) and the control of the textural properties of cooked rice (for example, soft, hard, and chewy).

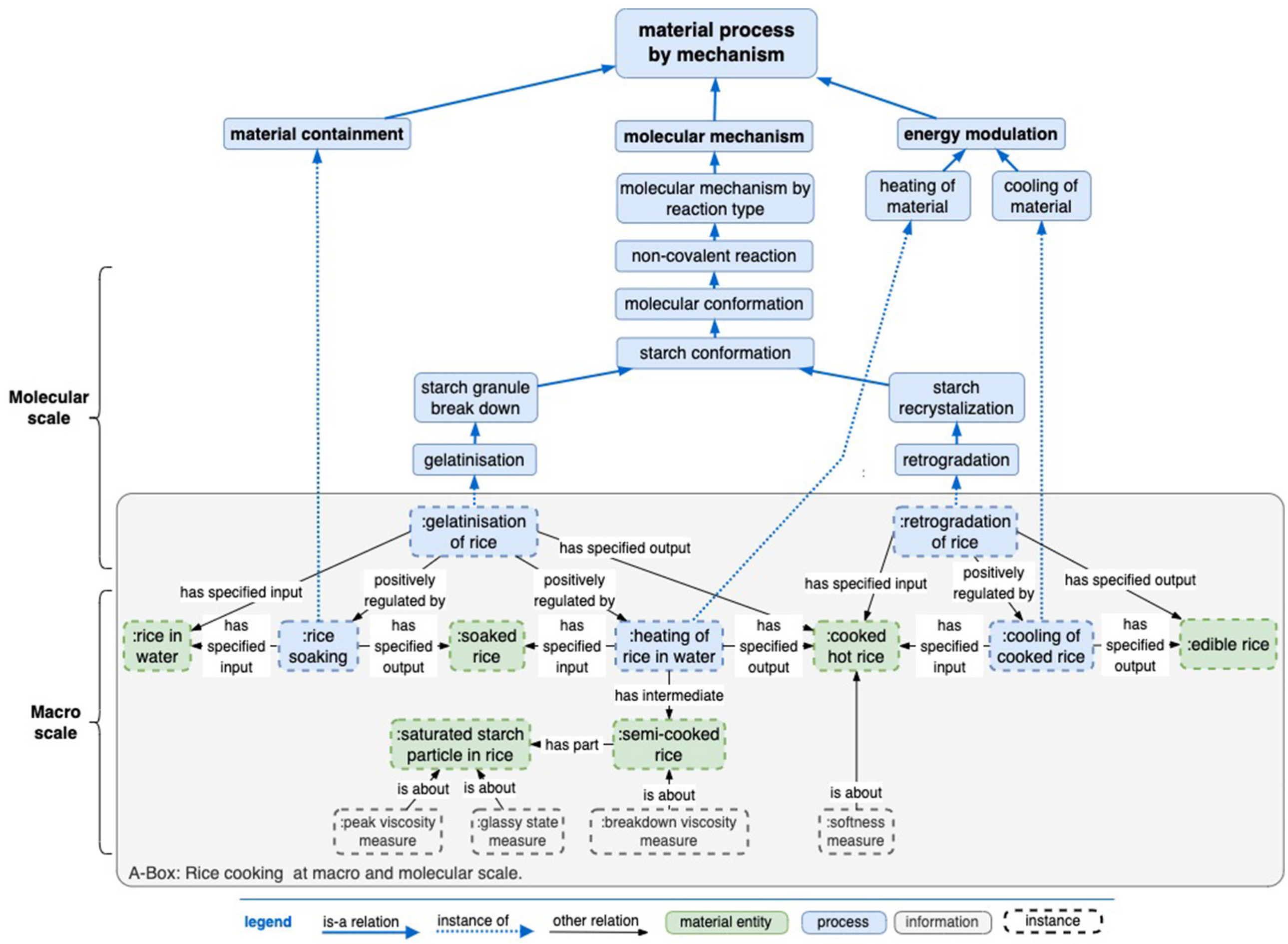


FIGURE 8
An example of a rice cooking process model can be viewed from both a macro and molecular scale, with an RO "positively regulated by" object property tying the two together.

Generally, ontology provides a formal language and framework to align various mechanisms, objectives, and instrumental measures. While the knowledge graph in this figure has been manually curated with food science expertise, this preliminary work could be evolved to support inference. Food science process terms such as peak viscosity and glass transition could also be text-mined from the literature and introduced under the ontology's mechanism and objective hierarchies at either a macro or a molecular scale.

The extracted terms can be used to structure data across food science experiments, and the analytical measurements associated with the process terms can be used for dynamic process control. Once finalized into ontologies such as OBI or FoodOn, this work will support data curation objectives—FAIR guidelines I1, R1.2, and R1.3 (FAIR Principles, 2017)—wherein data are coded in easily interpretable formats with precise provenance, and which use standardized (interoperable, reusable) language throughout. The chain of processes that ultimately generate data can be detailed as an instance of a protocol (whether experimental or operational), enabling a graph of a protocol's process, device, input, output, operator, and other contextual components—via ontology term and relation identifiers—to achieve disambiguation, comparability, and provenance of resulting datasets.

3. Future work and conclusion

This work should enhance clarity in finding a home for each food process under the matching mechanism/action or objective hierarchies. It should enable further research into how OWL logic can support the identification of equivalent processes for use in dynamic, versatile food processing pipelines. These elements are essential for enabling dynamic processing pipelines that can search and select from a library of processing components based on goal and/or resource constraints such as available tools or operators (mechanisms/actions) or material resources—a capability that humans often demonstrate in laboratory, industrial, or home food preparation settings. Additionally, this work should encourage the development of a better food processing protocol detail vocabulary, allowing appropriate comparison of data points within food composition databases and nutritional studies.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

DD was primarily supported by the USDA Non-Assistance Cooperative Agreement 58-8040-8-014-F and Genome Canada Grant 286GET.

Acknowledgments

The authors would like to acknowledge colleagues Magalie Weber of INRAE and Rhiannon Cameron of CIDGOH for their feedback on the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

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, e0154556. doi: 10.1371/journal.pone.0154556

Core Ontology for Biology and Biomedicine (n.d.). Available online at: https://obofoundry.org/COB/ (accessed October 15, 2022).

Doi, M. (2013). "Surfaces and surfactants," in *Soft Matter Physics*. Oxford: Oxford University Press. p. 51–71. doi: 10.1093/acprof:oso/9780199652952.003.0004

Dooley, D., Weber, M., Ibanescu, L., Lange, M., Chan, L., Soldatova, L., et al. (2022). Food process ontology requirements. *Semant. Web*. (2022) 22, 1–32. doi: 10.3233/SW-223096

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., et al. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, *NPJ Sci. Food*. 2, 23. doi: 10.1038/s41538-018-0032-6

FAIR Principles (2017). GO FAIR. Available online at: https://www.go-fair.org/fair-principles/ (accessed November 3, 2022).

FoodOn Robot Tables. (2023). *Google Docs*. Available online at: https://docs.google.com/spreadsheets/d/1VJtz4m67tdUNDqRe3m1Okdxll64nTR46GSvCOmb0APE/edit (accessed October 15, 2022).

Hinojosa-Nogueira, D., Pérez-Burillo, S., Navajas-Porras, B., Ortiz-Viso, B., de la Cueva, D. P., Lauria, F., et al. (2021). Development of an unified food composition database for the European project "Stance4Health," *Nutrients*. 13, 4206. doi: 10.3390/nu13124206

Jackson, R. C., Matentzoglu, N., Overton, J. A., Vita, R., Balhoff, J. P., Buttigieg, P. L., et al. (2021).BO Foundry in 2021: operationalizing open data principles to evaluate ontologies, *BioRxiv*. 06, 446587. doi: 10.1093/database/baab069

Kadam, S. U., Tiwari, B. K., and O'Donnell, C. P. (2015). "Improved thermal processing for food texture modification," in *Modifying Food Texture*, Chen, J., and Rosenthal, A. (eds.). Sawston: Woodhead Publishing. p. 115–131. doi: 10.1016/B978-1-78242-333-1.00006-1

Kulvatunyou, B., Drobnjakovic, M., Ameri, F., Will, C., and Smith, B. (2022). *The Industrial Ontologies Foundry (IOF) Core Ontology*. Tarbes: Formal Ontologies Meet Industry (FOMI). Available online at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=935068 (accessed June 25, 2023).

Naravane, T., and Lange, M. (2018). *Ontological Framework for Representation of Tractable Flavor: Food Phenotype, Sensation, Perception*. Available online at: http://ceur-ws.org/Vol-2285/ICBO_2018_paper_45.pdf (accessed October 16, 2022).

Naravane, T., and Tagkopoulos, I. (2023). Machine learning models to predict micronutrient profile in food after processing. *Curr. Res. Food. Sci.* 6, 100500. doi: 10.1016/j.crfs.2023.100500

Osorio, S., Scossa, F., and Fernie, A. R. (2013). Molecular regulation of fruit ripening. *Front. Plant Sci.* 4, 198. doi: 10.3389/fpls.2013.00198

Tao, K., Yu, W., Prakash, S., and Gilbert, R. G. (2020). Investigating cooked rice textural properties by instrumental measurements, *Food Sci. Human Wellness*. 9, 130–135. doi: 10.1016/j.fshw.2020.02.001

Temperatures and Chilling and Freezing Procedures. (2023). LII/Legal Information Institute. Available online at: https://www.law.cornell.edu/cfr/text/9/381.66 (accessed October 15, 2022).

Westenbrink, S., Presser, K., Roe, M., Ireland, J., and Finglas, P. (2021). Documentation of aggregated/compiled values in food composition databases; EuroFIR default to improve harmonization, *J. Food Compost. Anal.* 101, 103968. doi: 10.1016/j.jfca.2021.103968

Yamada, S. (2014). "Molecular interactions (molecular and surface forces)," in *Encyclopedia of Polymeric Nanomaterials*, eds S. Kobayashi and K. Müllen (Berlin; Heidelberg: Springer Berlin Heidelberg), 1–7.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 03 August 2023
DOI 10.3389/frai.2023.1191122

OPEN ACCESS

EDITED BY
Marie-Angélique Laporte,
Bioversity International, France

REVIEWED BY
Hai Wang,
China Agricultural University, China
Florence Amardeilh,
Elzeard, France

*CORRESPONDENCE
Baptiste Imbert
✉ baptiste.imbert@inrae.fr
Nadim Tayeh
✉ nadim.tayeh@inrae.fr

RECEIVED 21 March 2023
ACCEPTED 10 July 2023
PUBLISHED 03 August 2023

CITATION
Imbert B, Kreplak J, Flores R-G, Aubert G, Burstin J and Tayeh N (2023) Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes. *Front. Artif. Intell.* 6:1191122.
doi: 10.3389/frai.2023.1191122

Development of a knowledge graph framework to ease and empower translational approaches in plant research: a use-case on grain legumes

Baptiste Imbert[1]*, Jonathan Kreplak[1], Raphaël-Gauthier Flores[2,3], Grégoire Aubert[1], Judith Burstin[1] and Nadim Tayeh[1]*

[1]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne, Univ. Bourgogne Franche-Comté, Dijon, France, [2]Université Paris-Saclay, INRAE, URGI, Versailles, France, [3]Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, Versailles, France

While the continuing decline in genotyping and sequencing costs has largely benefited plant research, some key species for meeting the challenges of agriculture remain mostly understudied. As a result, heterogeneous datasets for different traits are available for a significant number of these species. As gene structures and functions are to some extent conserved through evolution, comparative genomics can be used to transfer available knowledge from one species to another. However, such a translational research approach is complex due to the multiplicity of data sources and the non-harmonized description of the data. Here, we provide two pipelines, referred to as structural and functional pipelines, to create a framework for a NoSQL graph-database (Neo4j) to integrate and query heterogeneous data from multiple species. We call this framework Orthology-driven knowledge base framework for translational research (Ortho_KB). The structural pipeline builds bridges across species based on orthology. The functional pipeline integrates biological information, including QTL, and RNA-sequencing datasets, and uses the backbone from the structural pipeline to connect orthologs in the database. Queries can be written using the Neo4j Cypher language and can, for instance, lead to identify genes controlling a common trait across species. To explore the possibilities offered by such a framework, we populated Ortho_KB to obtain OrthoLegKB, an instance dedicated to legumes. The proposed model was evaluated by studying the conservation of a flowering-promoting gene. Through a series of queries, we have demonstrated that our knowledge graph base provides an intuitive and powerful platform to support research and development programmes.

KEYWORDS

graph database, orthology, ontology, quantitative genetics, gene expression, comparative omics, Ortho_KB, OrthoLegKB

Introduction

To accelerate plant research and manage costs, model species first emerged as a good strategy for studying plant development and stress response, thus providing the research community with data and knowledge. Databases such as TAIR for *Arabidopsis thaliana* (Berardini et al., 2015), MTGD for *Medicago truncatula Gaertn.* (Krishnakumar et al., 2015), miyakogusa-jp for *Lotus japonicus* (Sato et al., 2008) or RAP-DB for *Oryza sativa* (Ohyanagi et al., 2006) were created to centralize, organize and curate the information on model species while providing tools for their analysis. Meanwhile, researchers working on other species have been inferring information from closely-related model plants using orthology and synteny. In fact, orthologs, i.e. genes descending from a common ancestral gene by a speciation event, are likely to have similar and conserved functions (Linard et al., 2021). However, it can prove difficult to identify the correct ortholog of a gene among its homologs based only on sequence similarity because of duplication events. Synteny and collinearity, i.e., conservation of the content and the order of genes on chromosomal regions, respectively, can help identifying orthologous blocks and hence deciphering true orthologous genes (Drillon et al., 2020). Such information is already made available and exploited through platforms such as PLAZA (Van Bel et al., 2012, 2022), sometimes supplemented by tools giving access to gene expression data (Kamei et al., 2016).

With the advent of new technologies, the once daunting sequencing costs have been dramatically reduced (Shendure et al., 2017), allowing for the production of high-quality assembled genomes including for orphan species (Ye and Fan, 2021). These new resources, along with associated annotations, are often being hosted on dedicated websites and/or made available in repositories of well-known databases such as NCBI (Sayers et al., 2022), Ensembl Plants (Yates et al., 2022), Gramene (Tello-Ruiz et al., 2021) or Phytozome (Goodstein et al., 2012). The release of the genome sequences is significantly boosting the production of genetic data to inform on the control of phenotypic traits by genes and the production of -omic data (mostly represented by genomics, transcriptomics, proteomics and metabolomics) to characterize and quantify the different molecules from a biological entity. However, efforts are still uneven across the broad spectrum of species since conducting experiments spanning a wide range of genotypes, tissues and conditions to generate solid data can be very informative, but also expensive and hard to achieve. Also, some quantitative trait loci (QTL) controlling quantitative traits still display low resolution, either due to low marker density or to low recombination rate in the respective genomic regions, which can result in large number of genes within the confidence intervals and long lists of candidate genes. Comparing QTL positions across species can help pinpointing orthologous ones and thus refining the intervals of those with low resolution. Such comparative translational research has also the potential to transfer functional information from one species to another or to a group of species.

Databases are powerful tools to leverage already produced datasets, not only as a mean of storage but also of intelligent exploitation. For example, the Comparative Genomics (CoGe) platform currently allows for the comparison of datasets from a wide range of organisms, with nearly 58,000 genomes available (Lyons and Freeling, 2008). Using sequence homology and synteny, researchers can identify structural and nucleotide variations for their species of interest. Researchers can also use the LoadExp+ extension to import experimental data in various common formats, such as VCF for polymorphism or FASTQ for RNA-seq, process them, and display the results as tracks in the genome browsers (Grover et al., 2017). Nevertheless, the data are predominantly stored using relational database management system (RDBMS), distributed in category-specific tables. One problem than can arise with RDBMS is that connecting tables containing large datasets during querying requires several joining operations, which are expensive in terms of time and computational resources (Vicknair et al., 2010).

The intuitive idea of structuring intertwined data into a graph was propelled by the World Wide Consortium for semantic web through the Resource Description Framework (RDF), (W3C, 1994). In order to obtain a logical model in RDF, each piece of data is sliced into atomic statements stored as triples, i.e., (1) the subject of the resource to describe, (2) a property assigned to the resource, termed predicate and (3) the object, either a description or another resource. The subject and the object are nodes in the graph, while the predicate is an edge connecting the two nodes (Abuoda et al., 2022). The directional decomposition of information allows the use of ontologies that organize knowledge and greatly improve data sharing in scientific communities (Stevens et al., 2020). However, databases using this format, called triplestores or RDF stores, are characterized by an atomic granularity of nodes which can make database modeling tedious. In addition, deep traversal of the graph requires self-joining of all traversed triples which can make the cost of traversing edges logarithmic (Donkers et al., 2020).

Alongside RDF, labeled-property graph (LPG) databases have emerged, currently led by Neo4j, which are fundamentally designed to improve graph traversal by directly storing on disk all existing edges between nodes. A benchmark from Khayatbashi et al. (2022) comparing RDF triple-stores and LPG databases with twelve queries shows that Neo4j is in fact more efficient to traverse multiple layers of data. Neo4j databases offer high flexibility by adding key-value properties to nodes and edges to effectively compact information, consequently making the modeling easier to read and to incrementally improve (Donkers et al., 2020; Neo4j, 2023b). Considering these assets, Neo4j databases were found advantageous to manage dense networks of information required for systems biology. The Reactome database (Fabregat et al., 2018) and its plant counterpart Plant Reactome (Naithani et al., 2019) have already switched from an RDBMS database to a Neo4j database, since metabolic pathways are intrinsically connected as a graph structure. In fact, using a graph database dropped the average query time of Reactome by 93 % (Fabregat et al., 2018). While a graph is intuitive when representing a biological pathway, the value of such modeling extends to many applications, including translational research. For instance, orthologous relationships across genes required for translational research, could be modeled in a Neo4j database with an "IS_ORTHOLOGOUS_TO" relationship between the two "Gene" nodes. Information regarding the gene identifier or annotation could be stored as internal node properties, available for querying. As the system is adaptable, new layers of data can successively be added and articulated. Omics Database Generator (ODG) is the first LPG designed for translational research as defined by Guhlin et al. (2017). ODG is a Neo4j graph-database, developed primarily for annotation transfer to non-model species of bacteria and plants. The structure of ODG has been made available for researchers to import their own data. Indeed, the comparison of newly generated data with existing data can confirm hypotheses or help to generate new ones. This is especially useful when datasets do not yield results supporting the initial research hypothesis, end up being set aside and remain unpublished (Raciti et al., 2018). It is therefore crucial to use as many available and high-quality datasets as possible, whether published or unpublished, as valuable sources of knowledge. However, ODG does not offer support for the integration of annotated genetic data, which is necessary for crop improvement, and it is likely to be difficult for non-expert users to understand its model and its underlying potential (Misra et al., 2019; Kaur et al., 2021).

The legume family (Leguminosae or Fabaceae) is the third largest family of flowering plants, with about 750 genera and nearly 19,500 species (The Legume Phylogeny Working Group et al., 2013). The Leguminosae include many taxa of agricultural or other economic importance and significant research efforts are needed to advance legume breeding and address the new challenges imposed to agriculture, namely production under climate change, with less pesticides and fertilizers. *Pisum sativum* L. (pea), *Lens culinaris* Medik. (lentil) and *Vicia faba* L. (faba bean) are examples of grain legumes that produce protein-rich seeds and play a key role in sustainable cropping systems (Guiguitant et al., 2020; Rubiales et al., 2021; Semba et al., 2021). Because of their large genomes, sometimes up to 30 times larger than the genome of the model legume *M. truncatula* (Jayakodi et al., 2023), the creation of -omics data on these species has lagged behind. In addition, data on a given species were mostly produced by the research community in the country of production, as the dominant production areas are sometimes different. Several databases have been developed that attempt to inventory the diversity of published datasets and provide tools to analyse and visualize them, including Soybase (Grant et al., 2010), the Pulse Crop Database (Humann et al., 2019), KnowPulse (Sanderson et al., 2019) and the Legume Information System (Berendzen et al., 2021). However, there is still a lack of options to link multi-species datasets together for further study.

LegumeIP is a relational database, initially created to transfer knowledge from model to crop legume species, and recently transformed into an integrative platform to support translational research, hosting homology, gene annotation and expression data for 17 legume species in its latest version (Li et al., 2012, 2016; Dai et al., 2021). Some recently sequenced cool-season legumes are however missing, including *P. sativum* (Kreplak et al., 2019), *L. culinaris* (Ramsay et al., 2021) and *V. faba* (Jayakodi et al., 2023). In addition, the interface of LegumeIP is designed to facilitate pairwise comparisons, from model species to less studied crop species, making the current design unsuitable for simultaneous comparison of multi-species experiments.

Here, we developed Ortho_KB, a robust framework for translational research in diploid plant species. We developed a first pipeline to compute homology and define syntenic chromosomal regions across species. This method was chosen to identify putative orthologs among homologs, thus establishing links between corresponding genes and connecting chromosomes. We designed a second pipeline to execute custom scripts that reformat all heterogeneous data files, including -omics datasets, for input into the database. Users can integrate published and unpublished information related to their species of interest including gene-phenotype associations from QTL data and expression information from transcriptomic resources and use the provided framework to get the most out of their data. Ortho_KB provides an intuitive database model that can be queried using Cypher language, to extract meaningful information in comma-separated values (CSV) files for further analysis. The framework has been applied to a subset of legume species, resulting in a database called OrthoLegKB, a multi-species and multi-omics graph-based database for collecting, integrating and querying heterogeneous data. OrthoLegKB currently allows the comparison of genetic, and -omic data from 5 legume species, i.e., *P. sativum*, *V. faba*, *L. culinaris*, *Vigna radiata (L.) R.Wilczek* and *M. truncatula*. Finally, a use-case is described to demonstrate how the combination of quantitative genetics and expression data is possible in OrthoLegKB and can benefit translational research.

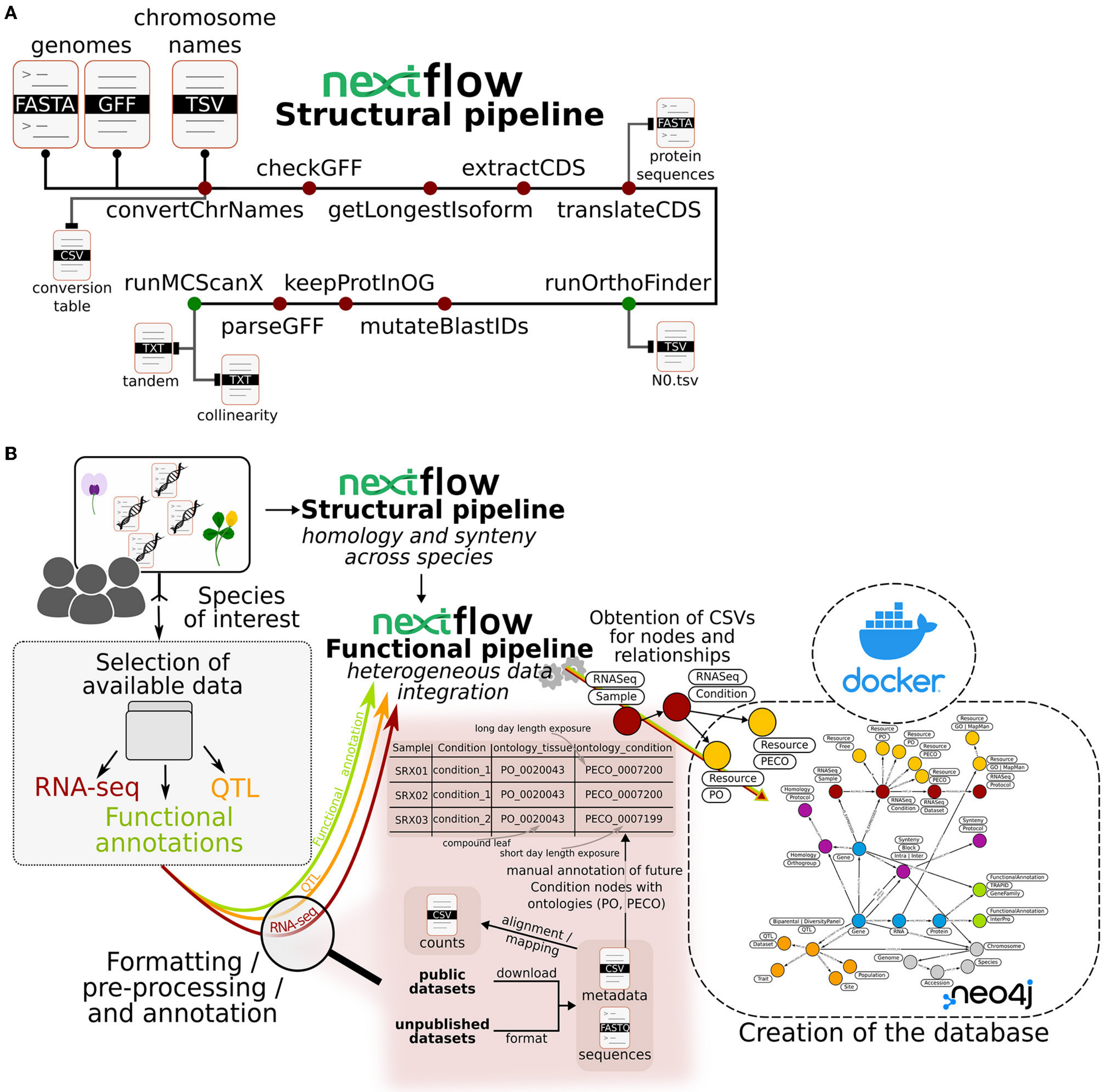
Materials and methods

Orthology and synteny

As illustrated in Figure 1A, in order to identify homologous genes and syntenic regions, genome FASTA and annotation files as well as an optional conversion table for chromosomes are used as input files. The conversion table must include the original chromosome ID and the desired ID in the database. Unique chromosome and scaffold IDs across species are more convenient for querying and are also required by synteny-visualization tools such as SynVisio (Bandi and Gutwin, 2020). The steps for synteny and orthology discovery are the following: (1) curate annotation files using the agat_convert_sp_gxf2gxf.pl parser from agat v0.9.1 by automatically removing duplicated features and/or IDs, inferring missing IDs or parent features; (2) filter annotation files to keep only the longest isoform using the agat_sp_keep_longest_isoform.pl script; (3) extract coding DNA sequences (CDS) using the agat_sp_extract_sequences.pl script (Dainat et al., 2022); (4) generate protein sequences using the Seqkit v2.3.0 translation module (Shen et al., 2016); (5) submit protein sequences in FASTA format to OrthoFinder v2.5.4 with its default parameters using Diamond v2.0.12 in ultra_sensitive mode for the alignment instead of BLAST (Emms and Kelly, 2015, 2019; Buchfink et al., 2021). Finally, to connect homologous chromosomal regions, the OrthoFinder output is used to obtain syntenic blocks. First, alignment files are filtered to retain only pairs of proteins that are part of the same orthogroup. Second, these filtered alignment files are provided to MCScanX along with a merge of annotation files from all considered species (Wang et al., 2012). A minimum number of 10 genes to form a collinear block is set by default in the pipeline. All above-mentioned steps were included in a single pipeline, called the structural pipeline, using Nextflow (Di Tommaso et al., 2017).

Functional gene annotation

Functional annotations were conducted by manually submitting CDS sequences to annotation tools either available on online platforms or to be run locally. The TRAPID online tool returned gene families, RNA families, and Gene Ontology (GO) terms associated with submitted genes (Bucchini et al., 2021).



Sample	Condition	ontology_tissue	ontology_condition
SRX01	condition_1	PO_0020043	PECO_0007200
SRX02	condition_1	PO_0020043	PECO_0007200
SRX03	condition_2	PO_0020043	PECO_0007199

FIGURE 1
Schematic representation of the pipelines used to build Ortho_KB, a NoSQL graph database framework for translational research. **(A)** The structural pipeline computing homology between genes and synteny across chromosomal regions from selected annotated genomes. All processes included in the pipeline, except those producing the mandatory final outputs, are represented by dark red circles. Processes producing the mandatory final outputs are represented by green circles. **(B)** General overview of the steps leading to the construction of an instance of Ortho_KB. Datasets that can be managed include RNA-seq data, QTL and functional annotations. As an example, we develop the treatment of an RNA-seq dataset from public or private origin. Alongside a regular extraction of counts, metadata of the samples must be annotated using ontologies to describe in particular the tissue of origin (Plant Ontology) and the experimental conditions to which the sample was subjected to (Plant Experimental Conditions Ontology). The functional pipeline will process inputed files and in this case the annotated metadata file will produce "Sample" and "Condition" nodes in the graph. This last node will also be connected by relationships to "Resource" nodes corresponding to the ontologies, thereby conserving the metadata information in the Neo4j graph database. The graph database is included in a Docker container, as shown on the right-hand side of the schema.

Genes that share sequence homology are gathered in gene families. "GeneFamily" nodes hold links to the Plaza website on which information regarding family-associated GO annotations and InterPro domains are available (Van Bel et al., 2022). To attribute summarized functions to genes, we assigned MapMan bins using the online Mercator4 (https://www.plabipd.de/portal/web/guest/mercator4) which resulted in a hierarchical annotation (Lohse et al., 2014; Schwacke et al., 2019). We also used eggNOG-mapper to obtain human-readable annotation and gene symbols from protein sequences (Huerta-Cepas et al., 2019; Cantalapiedra et al., 2021). A name is assigned to each gene when available in the literature. For instance, *MtrunA17Chr3g0135361* is annotated as *ELF3* for *EARLY FLOWERING 3*. Predicted proteins of each species were further annotated by locally running InterProScan v.5.53 with the "iprlookup" option (Jones et al., 2014; Blum et al., 2021), notably using databases such as Pfam (Mistry et al., 2021), Gene3D (Lees et al., 2012) or PANTHER (Mi and Thomas, 2009).

Genetic data extraction

The exact set of mandatory and optional information required to describe QTL data in Ortho_KB are described in the documentation available on the dedicated Git repository (see "Data availability statement" section). Briefly, the identifier, the trait name, and the associated markers are essential. A QTL arising from a study on a biparental mapping population is defined by a physical position on a chromosome between two flanking genetic markers and a peak marker within the confidence interval, if the information is available. A QTL from a genome-wide association study (GWAS) analysis is defined by a single marker location on a chromosome corresponding to the peak marker unless linkage disequilibrium data are available, then data is processed similarly to a QTL from a biparental population. Therefore, a QTL record might have information for one up to three markers. QTL data in the current version of OrthoLegKB were collected from published research articles (Supplementary Table S3).

Transcriptomic data extraction and expression quantifications

RNA-seq datasets were manually selected from NCBI. Sample IDs associated within a BioProject were collected in each case using esearch from entrez-direct v16.2 (Kans, 2013). The sample list was fed into nf-core/fetchngs pipeline v1.7 with the option "nf_core_pipeline rnaseq" to obtain all FASTQ files along with a metadata file (Patel et al., 2022). The nf-core/rnaseq v3.8 pipeline (Patel et al., 2023) was then run with the genome files, metadata file and FASTQ files with the arguments "skip_alignment", "pseudo_aligner salmon" and "salmon_quant_libtype A" to automatically assess strandedness (Patro et al., 2017). Salmon result files were finally processed into matrices for downstream analyses using tximport (Soneson et al., 2016). The "salmon.merged.gene_counts.tsv" file containing read counts and the "salmon.merged.gene_tpm.tsv" file containing the Transcript Per Million (TPM) normalized quantification were used for further processes. Samples listed in the metadata file originating from the nf-core/fetchngs pipeline were manually annotated to indicate the tissues used, the environmental conditions applied, and the experimental area (field, greenhouse, etc.), using the Plant Ontology (PO) and the Plant Experimental Condition Ontology (PECO) (Cooper et al., 2018).

Database construction and implementation

Graph database conceptualization

The current release of Ortho_KB was built as a NoSQL database framework to store and display data in a graph structure, using the Neo4j Community Edition v4.4.18 (Neo4j, 2023b).

We chose the Neo4j graph-database management system because of (1) its efficiency in handling highly connected data, (2) the graph algorithms already implemented and (3) the expressive Cypher query language it uses and (4) its capacity to import/export data using semantic web technologies. Entities, also called nodes, and edges, also referred to as relationships, were designed in a way to carry the biological information. Each gene or transcript is represented by a node and each gene is linked to its corresponding transcript by a relationship (e.g., gene *A* has a transcript RNA *A1*). Multiple properties can be stored and queried on nodes (e.g., RNA *A1* sequence length) and relationships (e.g., position of a protein domain on the protein sequence). In addition, one or more labels can be applied to nodes to group them into a set to facilitate querying. In this paper, labels are indicated by double quotation marks, for instance the "RNA" label for nodes of transcripts.

Input files processing

A Nextflow pipeline called functional pipeline, was created to process heterogeneous data from the previously described sources. The pipeline requires genome files, functional annotation files, RNA-seq files and QTL files to run. For the functional annotations, the pipeline includes a set of scripts to filter and format them into nodes and relationships, following the database model. For the GO annotations obtained from TRAPID, by default, only the most specific GO terms are retained for each gene by selecting those with parameter "is_hidden" equal to 0, resulting in a 90% reduction in the number of GO terms directly associated to genes. The GO W3C Web Ontology Language (OWL) file is downloaded and parsed to import the "is_a" and "part_of" predicates as relationships in the graph to allow graph traversal (W3C, 1994). Similarly, the provided annotation files from MapMan are used to create an ontology in TURTLE syntax using rdflib v.4.2.2 (Grimnes et al., 2023). For RNA-seq, salmon pseudo-counts are by default filtered to retain only genes for which the sum of TPM across samples is >5, to avoid creating many relationships for non-expressed genes. Gene expressions in all samples from the same condition are averaged, and both arithmetic and geometric means are stored on the edge between the "Gene" and the "Condition" nodes. For genetic data, previously formatted files are processed to identify genes included in the confidence interval of QTL using pybedtools v.0.9.0 (Quinlan and Hall, 2010; Dale et al., 2011).

Briefly, for all processes, the pipeline creates CSV files to populate the database and a summary file listing all CSV files to be imported in a format readable by Neo4j. All nodes and relationships that can be generated are described in Supplementary Tables S2, S3.

Database implementation

A Bash script was written to create and populate the database. It includes three steps. The first step prepares the import environment by building a Docker container. Running the

Docker container will start the database, by default available at http://0.0.0.0:7474/browser/. The second step performs the import to populate the Neo4j database using the neo4j-admin import command. The third step imports the PO, PECO, GO and MapMan ontologies using the n10s.onto.import.fetch method from the neosemantics (n10s) plugin (Barrasa, 2022). The import creates a node per term, connected to broader terms by a "SCO" relationship obtained from the property rdf:subClassOf. The nodes of the resulting subgraph are then labeled according to their source (PO, PECO etc.) and connected to the rest of the graph using a set of Cypher queries.

Plant species selection for OrthoLegKB

For this study, five species were chosen. These include the model legume *M. truncatula*, three cool-season legumes of agronomic importance, i.e., *P. sativum*, *L. culinaris*, *V. faba,* and a relatively distant warm-season legume species, *V. radiata*. All species belong to the Galegoids subclade, with the exception of *V. radiata*, which is part of the sister group, the Milletoids subclade. We selected the latest genomic data from *P. sativum* cultivar Cameor v.1 assembly (Kreplak et al., 2019), *M. truncatula* accession A17 v.5 assembly and v.1.9 annotation (Pecrix et al., 2018), *V. faba* accession Hedin/2 v.1.0 assembly (Jayakodi et al., 2023), *L. culinaris* cultivar CDC Redberry v.2.0 assembly (Ramsay et al., 2021) and *V. radiata subsp. radiata* cultivar VC1973A v.6 assembly (Ha et al., 2021). All genomes were assembled into chromosomes, generated using long-read technology, except for *P. sativum*. The *M. truncatula* annotation file was filtered to keep only features from EuGene and BioFileConverter. Gene prefixes were also modified using a custom script. Details on selected genome assemblies and genome statistics are available in Table 1.

Data visualization

The visualization of the graph model was created using Arrows (Neo4j, 2023a). The UpSet plot was created using UpSetR (Conway et al., 2017). Visualization of large-scale synteny was performed with the SynVisio online tool (Bandi and Gutwin, 2020) or with tailored R scripts, while microsynteny was plotted using the R package gggenomes (Hackl and Ankenbrand, 2023).

Hardware and query time

The server hosting OrthoLegKB is based on an OpenStack infrastructure, with 4 virtual CPUs and 8 Gb of RAM. For each query presented in the Results section, the average response time over five iterations was indicated.

Results

Ortho_KB is a framework for translational research in plant species

Studying a particular trait or gene often requires the collection of different types of information available on different websites and databases, for the species of interest as well as for close species. We have created Ortho_KB, a database framework built with successive pipelines to facilitate the exploration of all data relevant to a trait or gene of interest in a single environment. Ortho_KB provides a unique and multi-functional structure that can be populated with datasets of interest and then queried for comparative and functional genomics studies. The current Ortho_KB modeling aims at enabling translational research across a wide range of selected species by making data easily searchable and the process more straightforward. The framework relies heavily on orthology and synteny relationships to build bridges between species, and transfer and/or compare genetic and genomic information between them. A Nextflow pipeline, called the structural pipeline (Figure 1A), first identifies groups of homologous genes – orthogroups – based on protein sequence similarity. It then looks for conserved gene order between pairs of chromosomes, within or between species, to highlight collinear regions. Homologs in collinear regions are more likely to be orthologs and therefore have similar functions. A second Nextflow pipeline, called the functional pipeline, connects information from the first pipeline and additional data available from separate tables including gene annotation, gene expression and QTL positions (Figure 1B). All heterogeneous data are thus properly formatted for integration into the database.

Ortho_KB uses Neo4j graph-database management system

The Neo4j graph-database management system handles entities as nodes and their connections as relationships. In Ortho_KB, the data model revolves around "Gene" nodes, characterized by their start and stop positions on chromosomes (Figure 2). The "Gene" nodes are connected to their putative transcript ("RNA") nodes, themselves connected to the predicted proteins ("Protein") resulting from the translation of their RNA sequences. Homology and collinearity information computed using the structural pipeline create bridges across species at the gene and the chromosome levels, respectively. The current version of Ortho_KB includes 29 core categories of nodes tagged either by a single label, like "Gene" nodes or by a set of labels, like "RNASeq" supplemented by "Condition". They are connected by directed relationships, sometimes bearing additional properties (Figure 2). Individual nodes are defined by a unique identifier. For example, a "Gene" node is defined by a gene ID, matching the feature ID from the General Feature Format 3 (GFF3) annotation file, unique across species. Ortho_KB can be queried through the web Neo4j Browser, the terminal or other interfaces provided by Neo4j (Neo4j, 2023b).

Ortho_KB integrates different categories of data including gene annotation, genetic and transcriptomic resources

As shown in Figure 1, Ortho_KB gathers different categories of data.

In terms of functional annotation, complementary information sources are handled. These include TRAPID's gene families, GO annotations, MapMan bins and InterPro that are each integrated

TABLE 1 Specifications of species included in OrthoLegKB.

Species	Genotype	Assembly size (Mb)	Number of chromosomes	Protein coding genes	Assembly references
Lens culinaris	CDC Redberry	3,760	7	58,243	Ramsay et al., 2021
Medicago truncatula	A17	430	8	44,626	Pecrix et al., 2018
Pisum sativum	Cameor	3,920	7	46,905	Kreplak et al., 2019
Vicia faba	Hedin/2	11,900	6	34,221	Jayakodi et al., 2023
Vigna radiata	VC1973A	476	11	30,882	Ha et al., 2021

Values for "Protein coding genes" take into account a single isoform per gene.

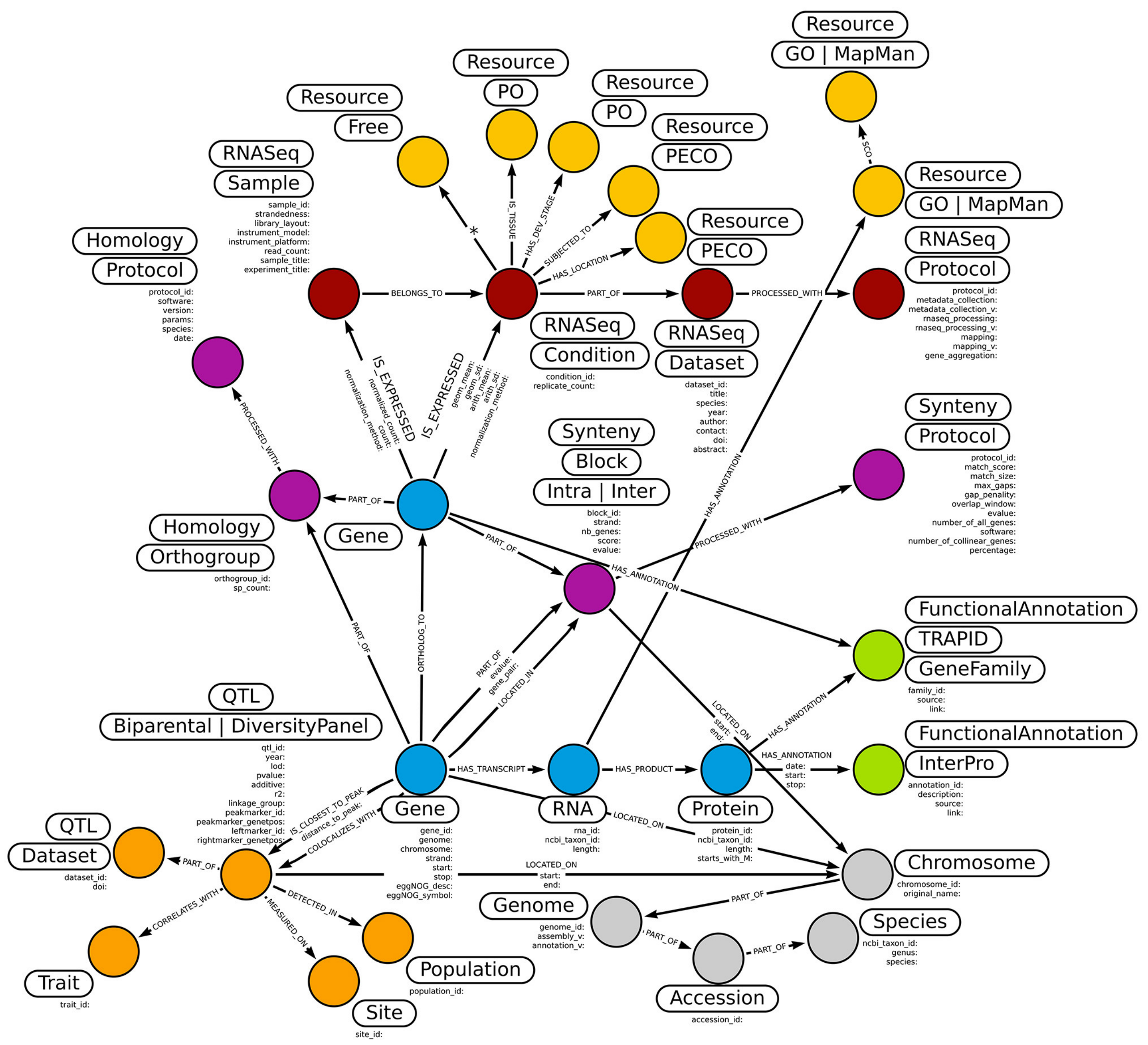


FIGURE 2
Overview of the Ortho_KB translational database model. In the graph model, colored circles represent the 29 core node types, which are entities with labels and properties. "Gene", "RNA", and "Protein" and related genomic nodes are shown in blue, "Homology" and "Synteny" and related nodes in mauve, ontology term nodes in yellow, the RNA-seq nodes in dark red, functional annotation nodes in light green, taxonomic nodes in light gray, and QTL-related nodes in orange. The category of each node is described by the associated labels, which are contained in elongated boxes near the nodes, and the properties correspond to the lists of elements placed below the labels. Nodes are connected to each other by relationships, represented by arrows, which can also store information as properties.

in a separate node type. TRAPID gene families and Mapman bins provide synthetic overviews of gene functions while GO annotations and InterPro provide detailed descriptions focusing on gene functions and protein domains, respectively (Figure 2).

Regarding genetic data, the model includes connections between genes and QTL information either resulting from QTL mapping in biparental populations or GWAS in diversity panels. Since the two mapping approaches are grouped with the "QTL" label, we added a second label, either "BiparentalPopulation" or "DiversityPanel" to differentiate them. All genes located within the confidence interval of a QTL are connected to the "QTL" node with a "COLOCALIZES_WITH" relationship. The closest gene to the peak marker is additionally connected to the "QTL" node with a "IS_CLOSEST_TO_PEAK" relationship with its distance to the peak marker included as a property. Additional information to describe a QTL are included in connected nodes such as the experimental geographical "Site", the studied "Population" and the "Trait" (Figure 2).

For transcriptomics, we have developed scripts to handle read counts. Read counts can either be generated using the pipeline of Patel et al. (2023) to optimize comparability of data (see "Materials and methods" section), or according to the method chosen by the user. The user is also free to integrate data previously analyzed with other methods. Replicates originating from the same biological condition are summarized into a condition that has to be manually annotated with ontology terms describing best the experimental conditions and biological material. The PECO and PO ontologies were selected for this purpose (Cooper et al., 2018). Using n10s inference, this model allows to traverse the ontology and unveil datasets from experiments performed in similar conditions (Barrasa, 2022). If no ontologies are available to appropriately describe conditions, free terms might be introduced (Figure 2).

OrthoLegKB was developed with Ortho_KB to provide a translational tool for grain legumes

To prove how Ortho_KB can serve translational approaches and the research goals of a scientific community, we chose to apply it to five diploid legume species belonging to the Galegoid (cool-season legumes) and Milletoid (warm-season legumes) clades creating the OrthoLegKB database.

To leverage data from all five species using comparative genomics, we started by searching for orthologs with the structural pipeline using the latest genome assemblies. The pipeline was run for 740 CPU hours with 20 CPUs allocated (7 h 40 in real time), with a maximum physical memory usage of 46 Gb. In total, 14,565 out of 29,428 total orthogroups (49.49 %) were shared by all five species and 8.24 % by all species excluding *V. radiata,* the only representative of the Milletoid clade (Figure 3).

Then, public datasets with QTL and RNA-Seq data were mined, annotated with the ontologies used in Ortho_KB and included in the database using the functional pipeline. The pipeline was run for 11 CPU hours (14 min in real time), with a maximum physical memory usage of 2 Gb. A list of these datasets is available in Supplementary Table S1. OrthoLegKB currently contains more than 815,000 nodes and close to 15,000,000 relationships associated to the different types of data. The exact number of nodes in each category can be found in Supplementary Table S4.

OrthoLegKB can be used to address various scientific questions including the conservation of the control of flowering time in legumes

As a use-case to demonstrate how to exploit OrthoLegKB, we searched for the orthologs of a previously-studied flowering time regulator, the *FLOWERING LOCUS T* from *M. truncatula* (*MtFTa1*) and sought evidences for potential conserved function across species. For this use-case, we have decided to work only on cool-season legume species. *MtFTa1* has been thoroughly studied (Hecht et al., 2007, 2011; Laurie et al., 2011; Cheng et al., 2021) and its physical position on *M. truncatula* chromosome 7 (Mt07) is known. It is identified as *Medtr7g084970* (Laurie et al., 2011; Cheng et al., 2021) or *MtrunA17Chr7_39606925_39618489* in the GFF3 of the Mt5.0 (r1.9) genome annotation version. The first step was the identification of candidate orthologous genes from *P. sativum*, *L. culinaris* and *V. faba*. Several candidates could be found across chromosomes through a single query in 11 ms (Figure 4). OrthoLegKB was then searched for syntenic blocks encompassing these candidate genes. Synteny between chromosome 3 from *P. sativum* (Ps03), chromosome 6 from *L. culinaris* (Lc06), chromosome 5 from *V. faba* (Vf05) and Mt07 at the *MtFTa1* locus was revealed highlighting the orthologs (Figure 5A). The syntenic blocks in *L. culinaris* and *V. faba* displayed each one orthologous *FTa1* gene, while two possible orthologous genes were detected in *P. sativum* namely *Psat3g090720* and *Psat3g090680* (Figure 5B). According to the conservation of protein length and domains' annotation information from the PANTHER database stored in OrthoLegKB, *Psat3g090720* seemed to be more similar to *MtFTa1* (Figure 6). In fact, *Psat3g090680* corresponds to *FTa2* described in Hecht et al. (2011). To examine any possible links with flowering control and thus function conservation, we searched for all QTL related to flowering contained in the previously identified syntenic blocks, allowing to also return QTL that did not include *FTa1* genes in their confidence intervals (Supplementary Table S5). As depicted in Figure 7, the query identified three QTL from Aguilar-Benitez et al. (2021) on Vf05, located at the same nucleotidic positions, that were linked to the number of days from the sowing until 50 % of the plants had visible open flowers (DF50_09-10(2)_1) and the number of days from the sowing until the appearance of the first flower (DF1_07-08(3)_1 and DF1_06-07(2)_1). On Ps03, a QTL from Gali et al. (2018) corresponding to the number of days to flowering (PR15_26_1) was found upstream of the *FTa1* locus (2018). Two QTL from Williams et al. (2022), closer to the *P. sativum* locus and associated to the number of days to flowering (DTF3_1) and number of nodes on the main stem to the first flower in long days (DTF3_3) were also identified. The *L. culinaris FTa1* gene was the only gene to be part of the confidence interval of a flowering-related QTL, qDTF.6-2_1. qDTF.6-2_1 is a number of days to flowering QTL from Haile et al. (2021), and close to the qDTFL-6A_1 from Yuan et al. (2021) related to the number of

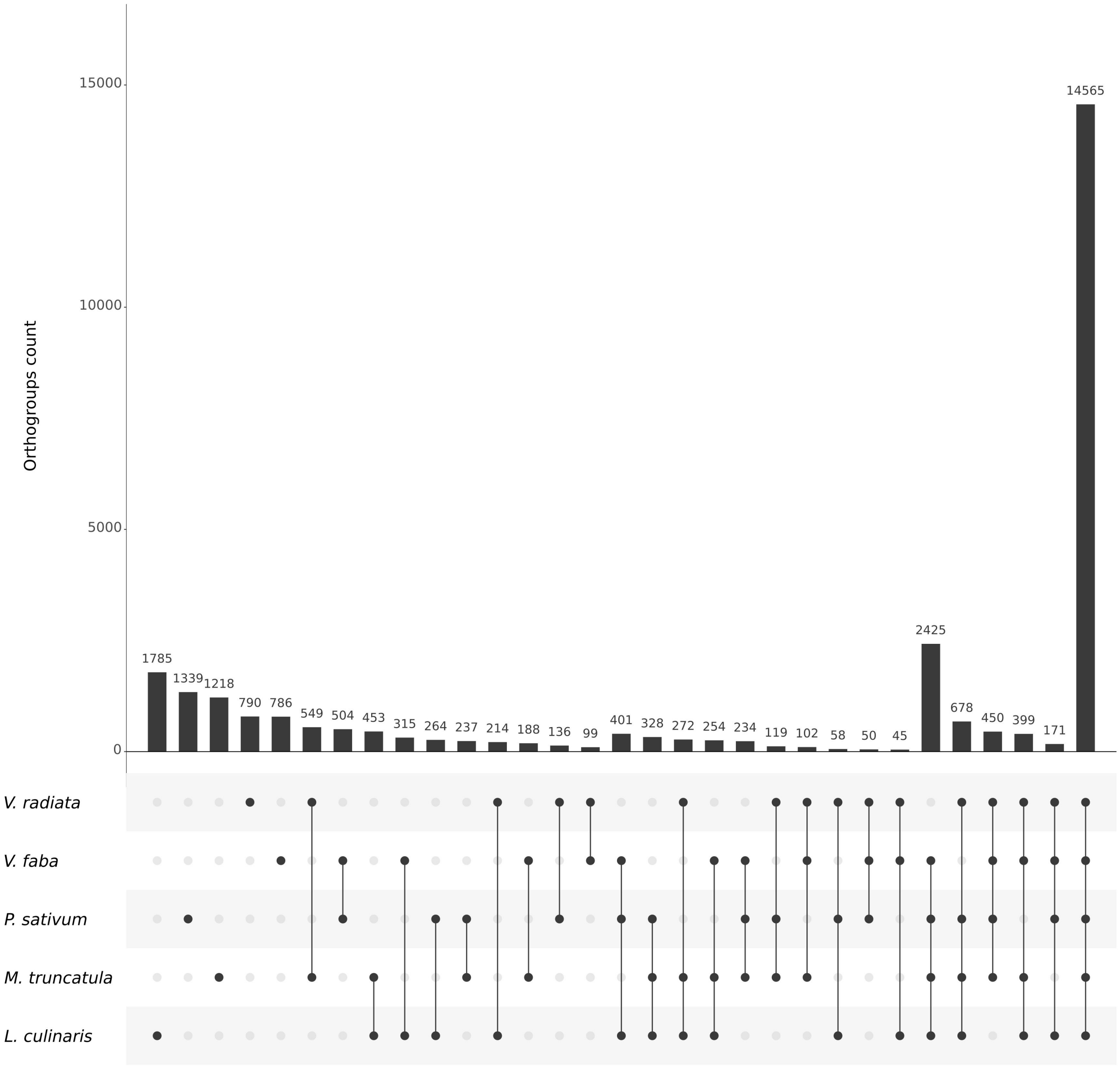
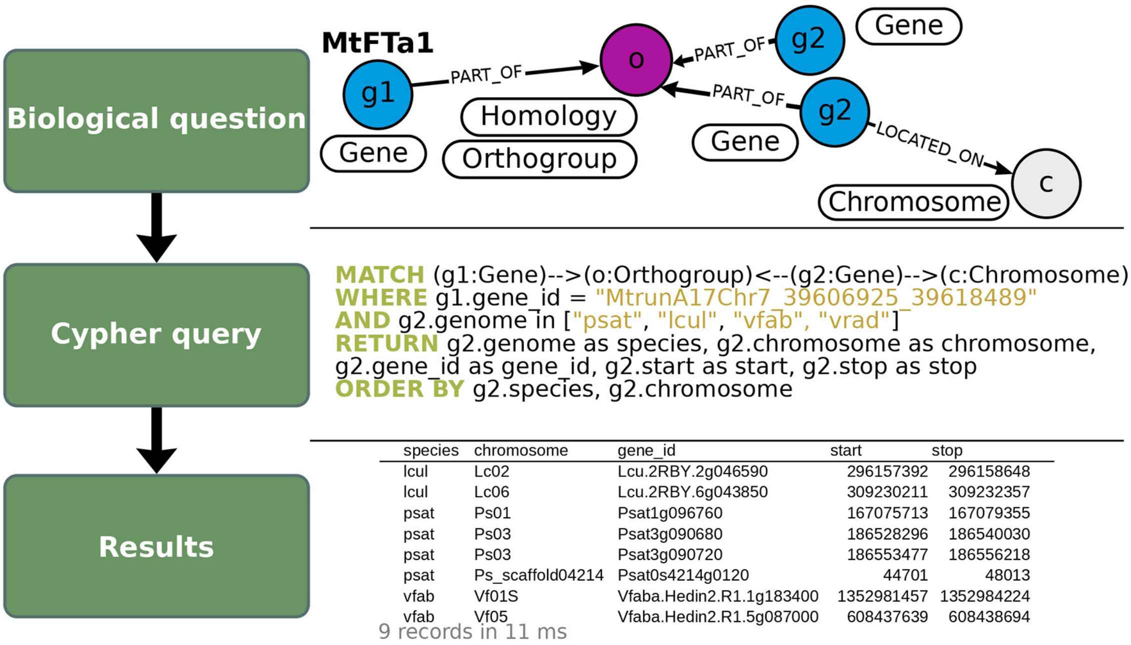


FIGURE 3
UpSet plot highlighting the number of orthogroups within and between legume species included in OrthoLegKB. The structural pipeline of Ortho_KB was used to identify the orthogroups. The bar plot shows the number of orthogroups for each possible set of species. The dots indicate the species associated with each bar.

days to flowering under low red/far-red light quality. Regarding expression, the *MtFTa1* gene is known to be mainly expressed in leaves and stems in *M. truncatula* (Laurie et al., 2011; Thomson et al., 2019). Therefore, we sought to investigate the top three tissues from the shoot system in which its orthologs were mostly expressed. Thanks to the inference allowed by the annotation of conditions with ontologies, we show that in the collection of RNA-seq datasets available for *M. truncatula* in OrthoLegKB, *MtrunA17Chr7_39606925_39618489* was mainly expressed in vegetative shoot apex, reproductive shoot apex and vegetative shoot system. *Psat3g090720* was expressed in the peduncle, stem and leaf tendrils. Transcripts from *Vfaba.Hedin2.R1.5g087000* were predominantly detected in adult vascular leaves, pods, and stems. For *L. culinaris*, all the experiments integrated were performed on leaves, in which the expression of *Lcu.2RBY.6g043850* was detected (Figure 8) and more particularly under far-red light conditions.

Discussion

This paper presented the use of knowledge graphs to integrate genetic and -omics data with the aim of facilitating translational research. The main philosophy was to provide a single environment where heterogeneous datasets from multiple species can be



```
MATCH (g1:Gene)-->(o:Orthogroup)<--(g2:Gene)-->(c:Chromosome)
WHERE g1.gene_id = "MtrunA17Chr7_39606925_39618489"
AND g2.genome in ["psat", "lcul", "vfab", "vrad"]
RETURN g2.genome as species, g2.chromosome as chromosome,
g2.gene_id as gene_id, g2.start as start, g2.stop as stop
ORDER BY g2.species, g2.chromosome
```

```
species  chromosome        gene_id                           start        stop
lcul     Lc02              Lcu.2RBY.2g046590             296157392   296158648
lcul     Lc06              Lcu.2RBY.6g043850             309230211   309232357
psat     Ps01              Psat1g096760                  167075713   167079355
psat     Ps03              Psat3g090680                  186528296   186540030
psat     Ps03              Psat3g090720                  186553477   186556218
psat     Ps_scaffold04214  Psat0s4214g0120                   44701       48013
vfab     Vf01S             Vfaba.Hedin2.R1.1g183400     1352981457  1352984224
vfab     Vf05              Vfaba.Hedin2.R1.5g087000      608437639   608438694
9 records in 11 ms
```

FIGURE 4
Illustration of the query used to search for putative orthologs of *MtFTa1* in OrthoLegKB. Putative orthologs in pea (psat), lentil (lcul), faba bean (vfab) and mung bean (vrad; **top panel**) were queried in Cypher **(middle panel)**, and several properties were returned in CSV format **(bottom panel)**. Genes belonging to the same orthogroup as *MtFTa1* were selected and their positions on the respective chromosomes were returned. Note that relationships' names were not displayed in the query section to keep it concise but were specified when running the query. The number of records returned in the output table and the average response time of the query are shown in light gray below the table.

accessed and examined with quasi-instantaneous querying time, thus allowing to address relevant biological questions, generate hypotheses, and transfer information from a single or group of species to others. The current version of the framework handles genome annotations, QTL and transcriptomic data. Users can identify orthologs, highlight candidate genes for specific traits, pinpoint possible pleiotropy and reveal conserved functional synteny. Ortho_KB gives the opportunity to capitalize on both published and unpublished datasets for further valorisation. The interest of such a database was demonstrated by populating Ortho_KB to create OrthoLegKB, a database dedicated to research on legume crop species, and supported by a use-case study focusing on a flowering-time gene.

Ortho_KB leverages recent analytical workflows and ontology standards to host high-quality data and ensure comparability across datasets

The Ortho_KB framework was built with the hypothesis that homologous genes found in collinear regions are most likely to be orthologs. Collinearity mitigates the effects of genome duplication and fractionation and thus most likely pinpoints true orthologs (Tang et al., 2008). Besides bridges between genomes based on orthology, additional information layers were incrementally integrated and connected to gene entities, taking advantage of the modeling flexibility allowed by Neo4j. The integration of such information was planned following homogenization rules for quality purposes. For expression data, we chose to use a single pipeline to process all transcriptomic datasets and avoid prejudice related to discrepancies in bioinformatic analysis protocols including alignment procedure, GC bias treatment. A similar initiative was taken for the gene atlas dedicated to *M. truncatula* (Carrere et al., 2021). We further decided to integrate normalized expression but not differential expression (DE). In fact, since the aim with Ortho_KB is to explore gene expressions across multiple samples and experiments, including expression in the form of DE would restrict analyses to a specific imposed comparison. Yet, the support for differential expression might be provided in the near future. Several actively updated ontologies (PO, PECO) were further selected to best describe the various experimental conditions from which the transcriptomic data were obtained. Since it requires human expertise, the annotation of samples with ontologies remains manual in the current version of the framework.

Regarding QTL, and unlike trancriptomics data, the reprocessing approach in sake of comparability could not be established so far as the analysis requires access to metadata, which are often sparsely provided in the literature. However, as FAIR standards are gaining in popularity, a unified approach might be considered for genetic data analysis in an upcoming version (Wilkinson et al., 2016). To ensure that positions of QTL for similar traits can be compared within and between species, homogeneity in traits denominations is required. This constraint is difficult to meet as a trait can be measured or named differently. For example, the flowering time might be considered by some authors as the time

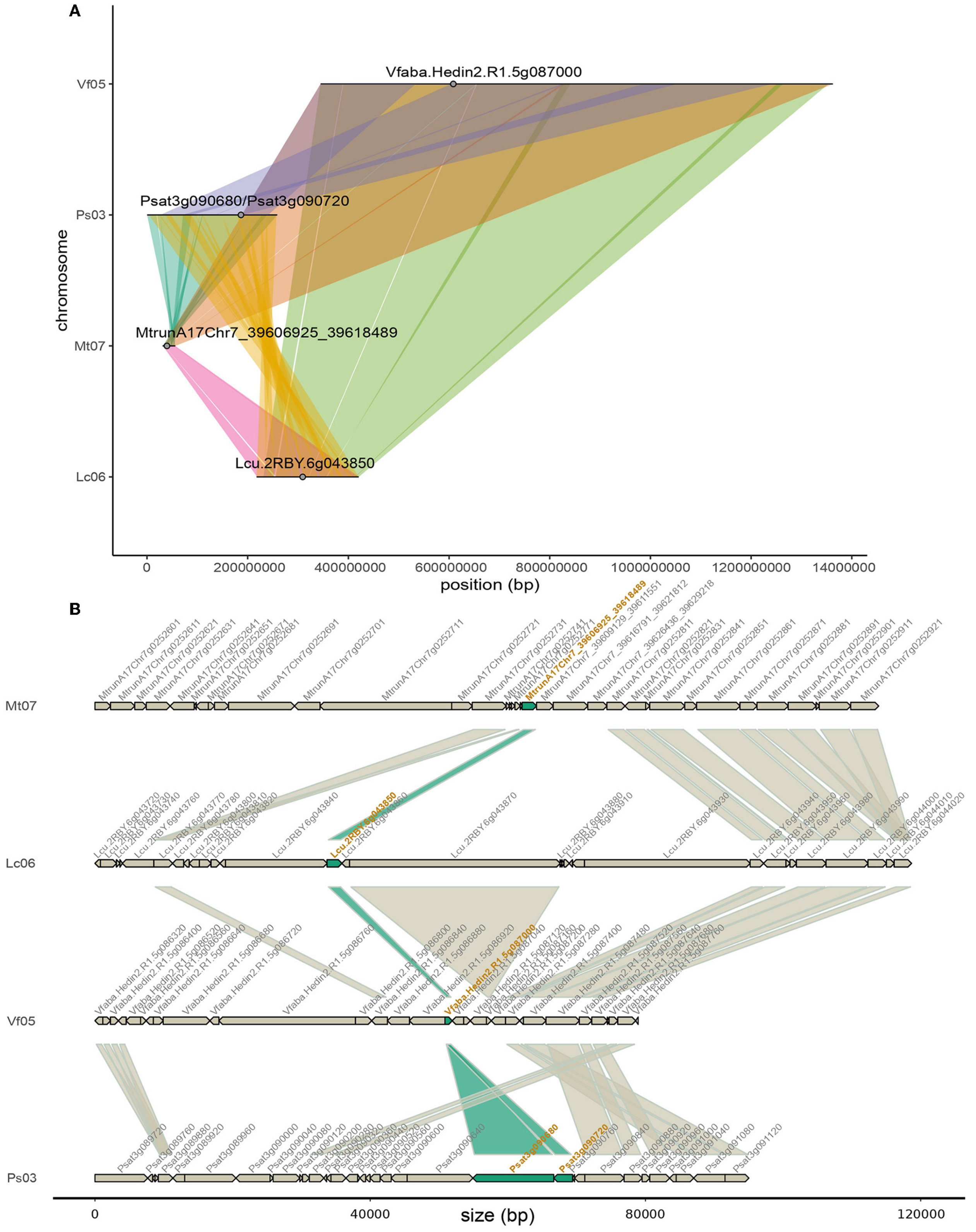
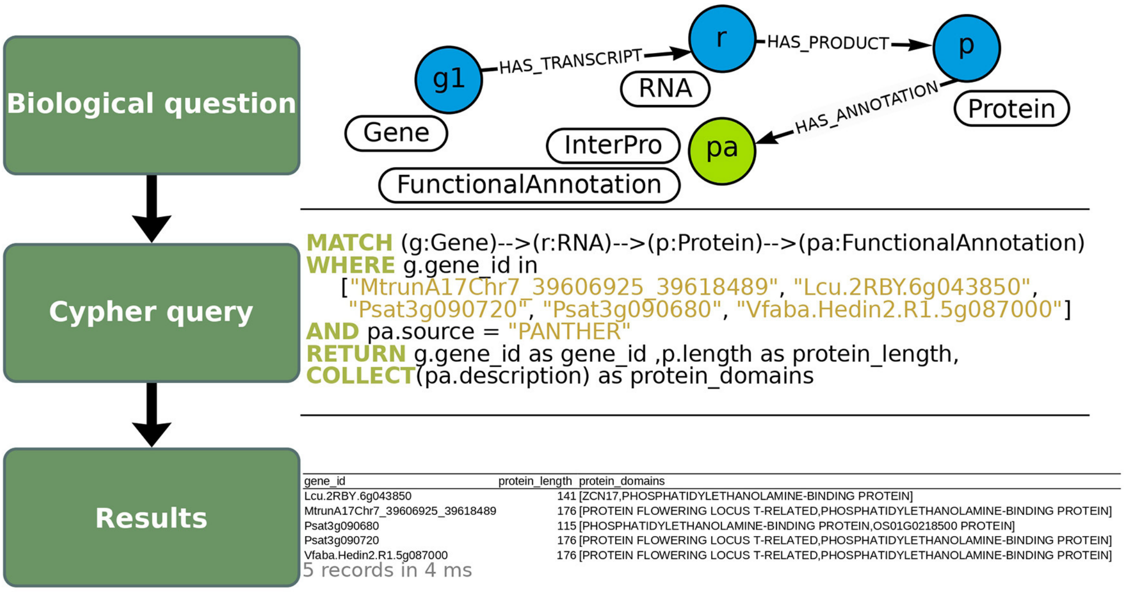


FIGURE 5

Macro- and micro-synteny of the chromosomal regions harboring *FTa1* or its orthologs in *M. truncatula, P. sativum, L. culinaris* and *V. faba*. **(A)** Macro-synteny at the chromosome level. *FTa1* and its orthologs are represented by gray dots on syntenic chromosome sections depicted as lines. Synteny between chromosomes is represented by ribbons. The positions of the two orthologs from *P. sativum* are shown even though they do not belong to any syntenic block in the database. **(B)** Micro-synteny of the *FTa1* loci. Genes are represented with arrows indicating the orientation of the open reading frames. Ribbons connect orthologous gene pairs. The IDs of *FTa1* orthologous genes are in orange and ribbons connecting them are filled in dark green. Since the four species have high genome size heterogeneity and variable intergenic sizes, intergenic regions were removed from the plot. Some gene names are not displayed due to space limitations. However, the gene sizes remain proportional.



```
MATCH (g:Gene)-->(r:RNA)-->(p:Protein)-->(pa:FunctionalAnnotation)
WHERE g.gene_id in
    ["MtrunA17Chr7_39606925_39618489", "Lcu.2RBY.6g043850",
     "Psat3g090720", "Psat3g090680", "Vfaba.Hedin2.R1.5g087000"]
AND pa.source = "PANTHER"
RETURN g.gene_id as gene_id ,p.length as protein_length,
COLLECT(pa.description) as protein_domains
```

```
gene_id                          protein_length  protein_domains
Lcu.2RBY.6g043850                           141 [ZCN17,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
MtrunA17Chr7_39606925_39618489              176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
Psat3g090680                                115 [PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN,OS01G0218500 PROTEIN]
Psat3g090720                                176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
Vfaba.Hedin2.R1.5g087000                    176 [PROTEIN FLOWERING LOCUS T-RELATED,PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN]
5 records in 4 ms
```

FIGURE 6
Extraction of protein domain annotations of *FTa1* and its orthologs using OrthoLegKB. "FunctionalAnnotation" nodes containing protein domain annotations **(top panel)** were queried in Cypher **(middle panel)**, for which several properties were returned in CSV format **(bottom panel)**. The nodes of protein domain annotations are connected to "Protein" nodes. Therefore, proteins corresponding to *FTa1* and its orthologous genes were selected, and their annotations from PANTHER were retrieved. Note that some relationships' names were not displayed in the query section to keep it concise but were specified when running the query. The number of records returned in the output table and the average response time of the query are shown in light gray below the table.

until the first plant has flowered, 50% of the plants have flowered or even 90% of them. Flowering time can also be expressed as the number of days between sowing and flowering or the number of degree-days. Nonetheless, a common vocabulary can be achieved with multi-species ontologies and needs to be developed. Such initiatives exist, such as the BBCH-scale framework to describe the phenological development stages of plants and serialized in RDF (Roussey, 2021), with instances for pea and faba bean but remain under-utilized. In the case of legumes, a higher-level ontology, not restricted to phenological stages, could use existing legume ontologies from the Crop Ontology, including the Lentil Ontology (CO_339) and the Faba bean Ontology (CO_365) (Shrestha et al., 2012). A general, consensus, ontology will however require manual work for the mapping of ontologies and its curation (Oellrich et al., 2015; Laporte et al., 2016; Cooper et al., 2018).

The graph model of Ortho_KB is intended to be regularly updated to enhance querying possibilities

The current version of Ortho_KB includes QTL and expression data but only allow the comparison of species based on single reference genomes. Lately, efforts on pangenomes and on the description of large diversity panels highlighted the importance of considering a wider set of accessions rather than a single representative one. As a first step toward the integration of structural variation, we intend to upgrade the graph model to allow hosting polymorphism variants in Ortho_KB. Since single nucleotide polymorphism (SNP) matrices constitute a large amount of data, the filtering and modeling will have to be thoroughly tested.

At the functional level, Ortho_KB presently provides solely transcriptomic data evidences. To provide complementary evidences regarding the function of genes of interest and their regulation at the post-transcriptional level, we plan to support the integration of proteomic data with Ortho_KB. This addition is also motivated by the ongoing standardization of proteomics output such as the mzTab format and downstream analyses (Griss et al., 2014; Ewels et al., 2020; Deutsch et al., 2022; Dubbelaar et al., 2022). Indeed, a recently published knowledge graph designed for clinical proteomic data namely the Clinical Knowledge Graph (CKG) accepts community-developed formats including mzTab and SDRF for metadata (Santos et al., 2022). Combining -omics layers can bring further evidence to a hypothesis and also open doors toward the understanding of complex underlying phenomena. Since, Ortho_KB was designed to be modular, one could even consider the inclusion of epigenomics information to gain insights on chromatin rearrangement during stress conditions for example. In this case, the integration of non-genic regions such as promoters and enhancers in the database could be evaluated.

Ortho_KB should be constantly evaluated to maintain performance and to facilitate its integration in the current databases ecosystem

As more biological data and data types are included, the Ortho_KB framework will have to be regularly fine-tuned to find

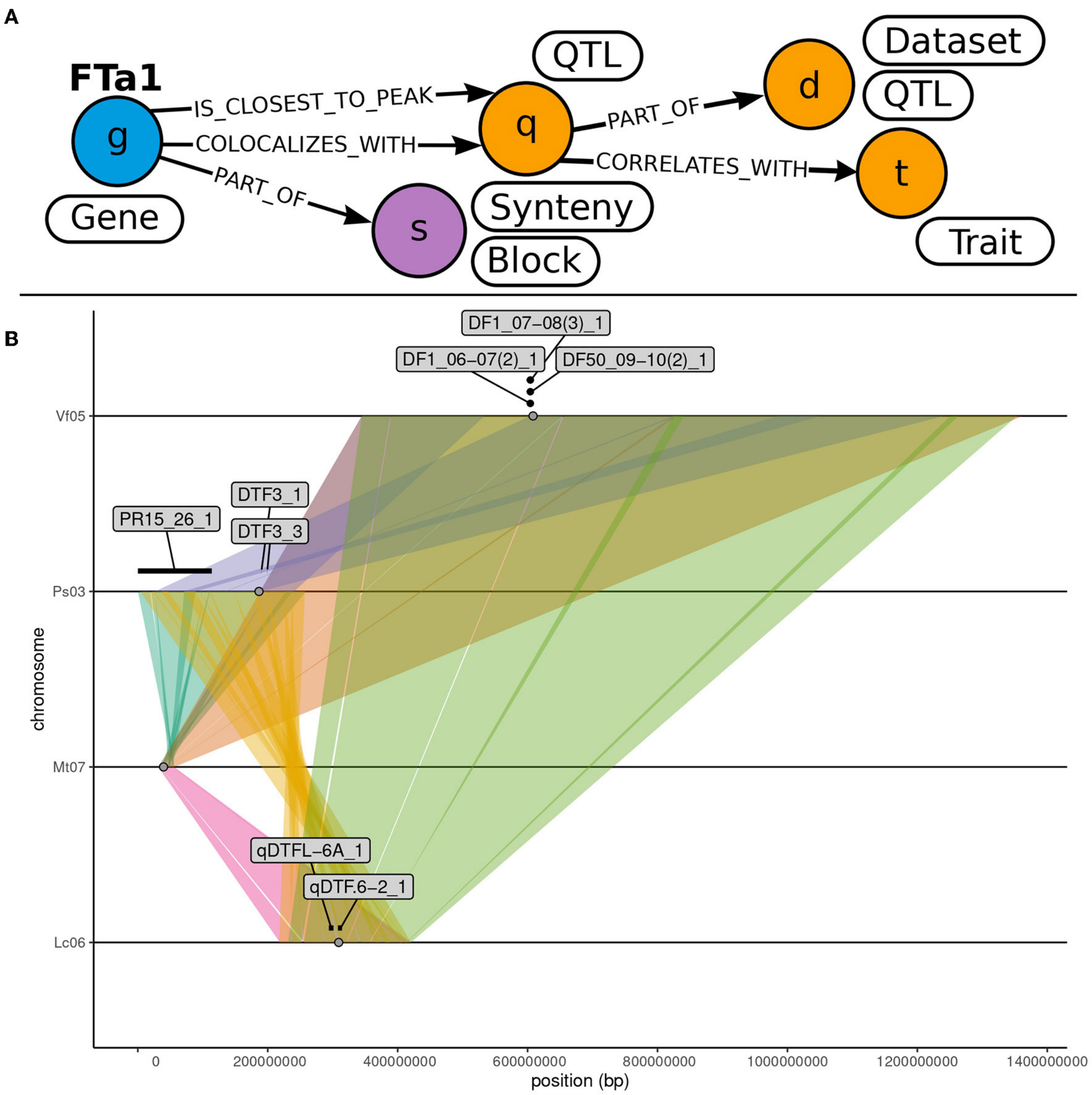
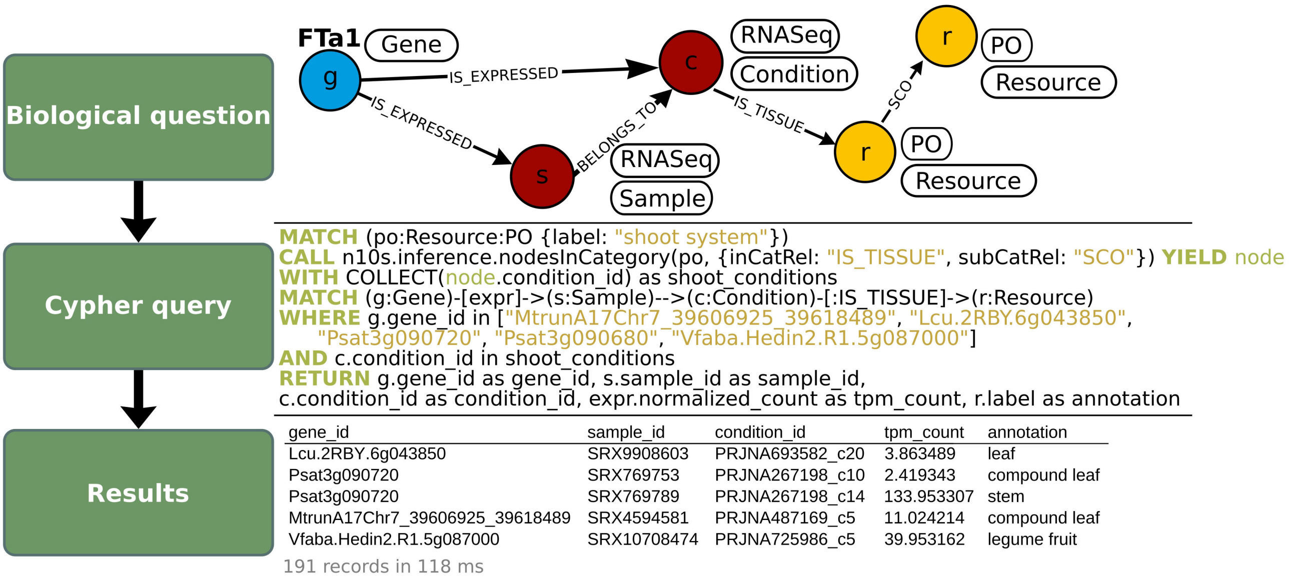


FIGURE 7
Identification of colocalising QTL with syntenic blocks hosting *MtFTa1* and its orthologs. **(A)** Illustration of the subgraph of OrthoLegKB queried to highlight QTL located near *FTa1* genes. "QTL" nodes contained within "Synteny" nodes including the *FTa1* gene were mined. Only QTL associated with flowering "Trait" were then kept. The query is available in Supplementary File S1. **(B)** Visualization of the colocalization between flowering QTL and syntenic blocks containing *FTa1* orthologs. Chromosome sections are represented by lines. Syntenic regions across chromosomes are represented by colored ribbons. *FTa1* and its orthologs are represented by gray dots. QTL labeled with their IDs are depicted by segments when information on both flanking markers is available or otherwise by simple dots.

the optimal graph model, but also in terms of the underlying configuration. In fact, for both the orthology backbone and the additional layers of the graph, single-property indexes have been created on properties that are regularly used as anchors to improve search performance at a small cost in storage space. Further guidance on the configuration of Neo4j has been previously published and will help to ensure high efficiency and scalability of Ortho_KB (Yoon et al., 2017). Several platforms already exist to study comparative genomics (Lyons and Freeling, 2008; Van Bel et al., 2022). The goal of Ortho_KB is different, since it mainly uses orthology and synteny as a way to transfer curated knowledge across species. Therefore, any created instance can be queried freely to answer complex tailored questions in a comprehensive manner.

As OrthoLegKB is primarily populated with published datasets, interoperability with already existing databases is essential. For RNA-seq, the NCBI Sequence Read Archive stores datasets according to defined rigorous standards (NCBI, 2023). QTL data, on the other hand, are typically scattered across multiple databases that store the information in different formats. Unlike the GWAS Catalog available for human (Sollis et al., 2023), no integrative databases store legumes QTL data in a unified format. Therefore, we plan to facilitate the integration of the content from existing legume databases. Other knowledge graph to understand the



```
MATCH (po:Resource:PO {label: "shoot system"})
CALL n10s.inference.nodesInCategory(po, {inCatRel: "IS_TISSUE", subCatRel: "SCO"}) YIELD node
WITH COLLECT(node.condition_id) as shoot_conditions
MATCH (g:Gene)-[expr]->(s:Sample)-->(c:Condition)-[:IS_TISSUE]->(r:Resource)
WHERE g.gene_id in ["MtrunA17Chr7_39606925_39618489", "Lcu.2RBY.6g043850",
    "Psat3g090720", "Psat3g090680", "Vfaba.Hedin2.R1.5g087000"]
AND c.condition_id in shoot_conditions
RETURN g.gene_id as gene_id, s.sample_id as sample_id,
c.condition_id as condition_id, expr.normalized_count as tpm_count, r.label as annotation
```

```
gene_id                         sample_id    condition_id        tpm_count    annotation
Lcu.2RBY.6g043850               SRX9908603   PRJNA693582_c20     3.863489     leaf
Psat3g090720                    SRX769753    PRJNA267198_c10     2.419343     compound leaf
Psat3g090720                    SRX769789    PRJNA267198_c14     133.953307   stem
MtrunA17Chr7_39606925_39618489  SRX4594581   PRJNA487169_c5      11.024214    compound leaf
Vfaba.Hedin2.R1.5g087000        SRX10708474  PRJNA725986_c5      39.953162    legume fruit
191 records in 118 ms
```

FIGURE 8
Fetch of expression levels (in TPM) of *MtFTa1* and its orthologs in different tissues of the shoot system, in OrthoLegKB. Normalized expression level of *FTa1* and its orthologs in RNA-seq samples **(top panel)** were queried in Cypher **(middle panel)**, and several properties were returned in CSV format **(bottom panel)**. The expression of *FTa1* genes was queried at the "Sample" level using the "expr" variable. The tissue annotations from the "Condition" nodes connected to these "Samples" were filtered to have only "Condition" nodes connected to the "PO" node "shoot system" or any of its more specific child terms. Note that in the table, the order of the rows was rearranged to show the diversity of annotations under the "shoot system" term. The number of records returned in the output table and the average response time of the query are shown in light gray below the table. The full table is available as Supplementary Table S6.

role of genes are already available. The KnetMiner software was created to analyse genome-scale knowledge graphs, with a recent support for the Cypher graph query language (Hassani-Pak et al., 2021). This platform allows to build gene networks based on semantics and information primarily extracted from the literature, including genetic data, phenotypes associated to SNPs or biological pathways. It was recently applied to wheat, generating networks for the *TT2* gene involved in pre-harvest sprouting (Hassani-Pak et al., 2021). In the specific case of legumes, the AgroLD triplestore is to our knowledge the only phenomics agronomy-centered database aiming at an integrative storage of biological information in the form of a knowledge graph (Venkatesan et al., 2018; Larmande and Todorov, 2021). Since Neo4j can handle RDF import and export, thanks to the neosemantics plugin, data exchange between OrthoLegKB and AgroLD could be considered to take advantage of both technologies. This goal is further supported by the ongoing development of the RDF-star extension which could support properties on edges of the graph (Abuoda et al., 2022). This would bridge the gap between LPG and RDF technologies for improved interoperability (Hartig, 2014). While the SPARQL RDF query language is common to all triplestores, Cypher from Neo4j is only used by the proprietary. However, the open-source GraphQL initiative known as GQL is seen as a potential technology agnostic standardization query language for graph databases (Donkers et al., 2020). We envisage that the legume research community will participate in the data collection and provide feedback on OrthoLegKB for regular improvement.

Ortho_KB offered an opportunity to develop a valuable tool for translational research in legumes, OrthoLegKB

We decided to select legume species to showcase how the Ortho_KB framework can serve translational research. OrthoLegKB is currently centered on few members mostly diploid cool-season legumes as the identification of orthologs is more straightforward than in polyploid species. Still, having a high-quality assembly is crucial for synteny detection and therefore true orthologs identification. The *FTa* locus in *P. sativum* (*PsFTa*) is in fact incorrectly assembled and annotated in the current version of the Cameor genome. While the *FTa1* gene we identified in *P. sativum* (*Psat3g090720*) is consistent with results from Hecht et al. (2011), the other copy (*Psat3g090680*) was reported as *FTa2* in the same study, both genes displaying similar expression patterns in leaves and apices, but with a weaker expression for *FTa2* (Hecht et al., 2011). In our study, the incomplete annotation of *Psat3g090680* in the version 1 assembly of *P. sativum* cv. Cameor most-likely prevented the creation of orthogroups correctly encompassing the *FT* gene family, and the subsequent inclusion of *PsFTa* in syntenic blocks. The locus displayed increased synteny with the other studied species, this time including *PsFTa1* and *PsFTa2* when considering the more recent *P. sativum* genome assembly from the Zw6 accession (Yang et al., 2022). A new assembly of the Cameor genome is expected soon and should improve the assembly and annotation of this region. Furthermore, supplementing OrthoLegKB with transcriptomic data will provide stronger support when searching for *FT* orthologous genes, by comparing their expression profiles. Fortunately, more legume genomes, have been lately assembled in high-quality using high-throughput chromosome conformation capture sequencing or long-read technologies, namely chickpea (Garg et al., 2022) or common vetch (Xi et al., 2022), which might reveal to be novel sources of data for OrthoLegKB. Thus, the graph will encompass more connected datasets, including information on abiotic and biotic stress response and be useful to a larger part of the legume research community.

The Ortho_KB framework is for the plant community and beyond

As demonstrated for legumes, the Ortho_KB framework is suitable for translational research within plant families to address common biological questions. Therefore, Ortho_KB could for instance be used in Solanaceae to study late blight attacking potato, tomato but not eggplants nor pepper. Genomes were sequenced for all these diploid species with long-reads technologies (Pham et al., 2020; Wei et al., 2020; Su et al., 2021; Liao et al., 2022). With more caution regarding the identification of orthologs, this resource would also meet the needs of research across plant families, or the needs of polyploids in the Brassicaceae and Poaceae families. Precise study of gene expression bias would be then crucial to identify expressologs (Das et al., 2016). While the scope of Ortho_KB was limited to plants for annotation reasons, its concept could be adapted for the benefit of other communities. For example, the wealth of draft-assembled diploid genomes profiting the Chelicerates community was recently exploited to highlight the conservation of chemosensory genes through comparative genomics (Vizueta et al., 2018). As more qualitative assemblies and associated -omics data are generated across plant and animal groups, we can only anticipate that the need for integrative multi-species databases will increase and Ortho_KB can contribute in this regard.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found below in the article/Supplementary material. OrthoLegKB with its user-guide is available for legume translational research at http://ortholegkb.versailles.inrae.fr/browser/. The functional pipeline for synteny is available at: https://forgemia.inra.fr/geapsi/pipeline/specifics_syntenymcscanx. The pipeline to create the translational database is available at: https://forgemia.inra.fr/geapsi/pipeline/specifics_ortho_kb. All scripts used to create the figures presented, including microsynteny from OrthoLegKB data, are available at: https://forgemia.inra.fr/geapsi/ecp-paper/ortholegkb_data.

Author contributions

BI, JK, and NT contributed to the conception and design of this work. BI was responsible for developing the pipelines, running the use-case, and writing the manuscript. JK and R-GF helped BI to build the graph database and provided bioinformatic support. GA collected information on the datasets to be included in OrthoLegKB and participated in the development of the use-case. JB provided ideas and managed funding acquisition. NT contributed to the scientific management of this work and was involved in drafting and writing the manuscript. All authors participated to manuscript revision, read, and approved the submitted version.

Funding

BI is the recipient of a PhD grant from French Ministry of Higher Education and Research. The work was part of the SPECIFICS project (ANR-20-PCPA-0008) funded by the French Priority Research Programme Growing and Protecting crops Differently (PPR-CPA), part of the national investment plan operated by the French National Research Agency (ANR).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1191122/full#supplementary-material

References

Abuoda, G., Dell'Aglio, D., Keen, A., and Hose, K. (2022). Transforming RDF-star to property graphs: A preliminary analysis of transformation approaches – extended version. *arXiv [Preprint]*. arXiv: 2210.05781. doi: 10.48550/arXiv.2210.05781

Aguilar-Benitez, D., Casimiro-Soriguer, I., Maalouf, F., and Torres, A. M. (2021). Linkage mapping and QTL analysis of flowering time in faba bean. *Sci. Rep.* 11, 13716. doi: 10.1038/s41598-021-92680-4

Bandi, V., and Gutwin, C. (2020). *Interactive Exploration of Genomic Conservation in Proceedings of Graphics Interface 2020 GI 2020*. Toronto: Canadian Human-Computer Communications Society/Société canadienne du dialogue humain-machine, 74–83.

Barrasa, J. (2022). *Neosemantics (n10s)*. Available online at: https://github.com/neo4j-labs/neosemantics (accessed December 21, 2022).

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome: tair: making and mining the "gold standard" plant genome. *Genesis* 53, 474–485. doi: 10.1002/dvg.22877

Berendzen, J., Brown, A. V., Cameron, C. T., Campbell, J. D., Cleary, A. M., Dash, S., et al. (2021). The legume information system and associated online genomic resources. *Legume Sci.* 3, 4. doi: 10.1002/leg3.74

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., and Kandasaamy, S., Mitchell, A., et al. (2021). The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., and Vandepoele, K. (2021). TRAPID 2.0: a web application for taxonomic and functional analysis of *de novo* transcriptomes. *Nucleic Acids Res.* 49, e101–e101. doi: 10.1093/nar/gkab565

Buchfink, B., Reuter, K., and Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18, 366–368. doi: 10.1038/s41592-021-01101-x

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293

Carrere, S., Verdier, J., and Gamas, P. (2021). MtExpress, a comprehensive and curated RNASEQ-based gene expression atlas for the model legume *Medicago truncatula*. *Plant Cell Physiol.* 62, 1494–1500. doi: 10.1093/pcp/pcab110

Cheng, X., Li, G., Krom, N., Tang, Y., and Wen, J. (2021). Genetic regulation of flowering time and inflorescence architecture by MtFDa and MtFTa1 in Medicago truncatula. *Plant Physiol.* 185, 18. doi: 10.1093/plphys/kiaa005

Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinf.* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364

Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., and Mungall, C., Sinn, B. T., et al. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.* 46, D1168–D1180. doi: 10.1093/nar/gkx1152

Dai, X., Zhuang, Z., Boschiero, C., Dong, Y., and Zhao, P. X. (2021). LegumeIP V3: from models to crops—an integrative gene discovery platform for translational genomics in legumes. *Nucleic Acids Res.* 49, D1472–D1479. doi: 10.1093/nar/gkaa976

Dainat, J., Hereñú, D., Davis, E., Crouch, K., LucileSol, F., Agostinho, N., et al. (2022). NBISweden/AGAT: AGAT-v1.0.0. *Zenodo*. doi: 10.5281/zenodo.7255559

Dale, R. K., Pedersen, B. S., and Quinlan, A. R. (2011). Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* 27, 3423–3424. doi: 10.1093/bioinformatics/btr539

Das, M., Haberer, G., Panda, A., Das Laha, S., Ghosh, T. C., Schäffner, A. R., et al. (2016). Expression pattern similarities support the prediction of orthologs retaining common functions after gene duplication events. *Plant Physiol.* 171, 2343–2357. doi: 10.1104/pp.15.01207

Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J., Mendoza, L., et al. (2022). The proteomexchange consortium at 10 years: 2023 update. *Nucleic Acids Res.* 5, gkac1040. doi: 10.1093/nar/gkac1040

Di Tommaso, D., Chatzou, P., Floden, M., Barja, E. W., and Palumbo, P. P. E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Donkers, A., Yang, D., and Baken, N. (2020). "Linked data for smart homes: comparing RDF and labeled property graphs," in *LDAC*. Available online at: https://ceur-ws.org/Vol-2636/02paper.pdf

Drillon, G., Champeimont, R., Oteri, F., Fischer, G., and Carbone, A. (2020). Phylogenetic reconstruction based on synteny block and gene adjacencies. *Mol. Biol. Evol.* 37, 2747–2762. doi: 10.1093/molbev/msaa114

Dubbelaar, M., Leon-Bichmann, Heumos, L., Peltzer, A., bot, nf-core, Scheid, J., et al. (2022). *nf-core/mhcquant: mhcquant 2.4.0 – Maroon Gold Boxer. Zenodo.* doi: 10.5281/zenodo.7389537

Emms, D. M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 16, 157. doi: 10.1186/s13059-015-0721-2

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., et al. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* 38, 276–278. doi: 10.1038/s41587-020-0439-x

Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., et al. (2018). Reactome graph database: efficient access to complex pathway data. *PLoS Comput. Biol.* 14, e1005968. doi: 10.1371/journal.pcbi.1005968

Gali, K. K., Liu, Y., Sindhu, A., Diapari, M., Shunmugam, A. S. K., Arganosa, G., et al. (2018). Construction of high-density linkage maps for mapping quantitative trait loci for multiple traits in field pea (Pisum sativum L.). *BMC Plant Biol.* 18, 172. doi: 10.1186/s12870-018-1368-4

Garg, V., Dudchenko, O., Wang, J., Khan, A. W., Gupta, S., Kaur, P., et al. (2022). Chromosome-length genome assemblies of six legume species provide insights into genome organization, evolution, and agronomic traits for crop improvement. *J. Adv. Res.* 42, 315–329. doi: 10.1016/j.jare.2021.10.009

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

Grimnes, G. A., Higgins, G., Hees, J., Aucamp, I., Arndt, N., Sommer, A., et al. (2023). RDFLib/rdflib: RDFlib 6.3.1. *Zenodo*. doi: 10.5281/zenodo.7748890

Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., et al. (2014). The mzTab Data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Prot.* 13, 2765–2775. doi: 10.1074/mcp.O113.036681

Grover, J. W., Bomhoff, M., Davey, S., Gregory, B. D., Mosher, R. A., Lyons, E., et al. (2017). CoGe LoadExp+: a web-based suite that integrates next-generation sequencing data analysis workflows and visualization. *Plant Direct* 1, 8. doi: 10.1002/pld3.8

Guhlin, J., Silverstein, K. A. T., Zhou, P., Tiffin, P., and Young, N. D. (2017). ODG: Omics database generator - a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. *BMC Bioinformatics* 18, 367. doi: 10.1186/s12859-017-1777-7

Guiguitant, J., Vile, D., Ghanem, M. E., Wery, J., and Marrou, H. (2020). Evaluation of pulse crops' functional diversity supporting food production. *Sci. Rep.* 10, 3416. doi: 10.1038/s41598-020-60166-4

Ha, J., Satyawan, D., Jeong, H., Lee, E., Cho, K., Kim, M. Y., et al. (2021). *A near-complete genome sequence of mungbean (Vigna radiata L.) provides key insights into the modern breeding program. Plant Genome.* 10, 121. doi: 10.1002/tpg2.20121

Hackl, T., and Ankenbrand, M. (2023). *Gggenomes: A Grammar of Graphics for Comparative Genomics*. Available online at: https://github.com/thackl/gggenomes (accessed March 20, 2023).

Haile, T. A., Stonehouse, R., Weller, J. L., and Bett, K. E. (2021). Genetic basis for lentil adaptation to summer cropping in northern temperate environments. *Plant Genome* 14, 144. doi: 10.1002/tpg2.20144

Hartig, O. (2014). *Reconciliation of RDF* and Property Graphs*. Available online at: http://arxiv.org/abs/1409.3288 (accessed March 13, 2023).

Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J. D., Amberkar, S., et al. (2021). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol. J.* 19, 1670–1678. doi: 10.1111/pbi.13583

Hecht, V., Knowles, C. L., Vander Schoor, J. K., Liew, L. C., Jones, S. E., Lambert, M. J. M., et al. (2007). Pea LATE BLOOMER1 Is a GIGANTEA ortholog with roles in photoperiodic flowering, deetiolation, and transcriptional regulation of circadian clock gene homologs. *Plant Physiol.* 144, 648–661. doi: 10.1104/pp.107.096818

Hecht, V., Laurie, R. E., Vander Schoor, J. K., Ridge, S., Knowles, C. L., Liew, L. C., et al. (2011). The Pea *GIGAS* gene is a flowering locus t homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. *The Plant Cell* 23, 147–161. doi: 10.1105/tpc.110.081042

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Humann, J., Jung, S., Cheng, C. H., Lee, T., and Zheng, P., Frank, M., et al. (2019). *A resource for pea, lentil, faba bean, and chickpea genetics, genomics and breeding. Proceedings of the International Plant and Animal Genome Conference*, 3. Available online at: https://www.pulsedb.org/

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednár, P., et al. (2023). The giant diploid faba genome unlocks variation in a global protein crop. *Nature* 26, 1–8. doi: 10.1038/s41586-023-05791-5

Jones, P., Binns, D., Chang, H. Y., Fraser, M., and Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Kamei, C. L. A., Severing, E. I., Dechesne, A., Furrer, H., Dolstra, O., Trindade, L. M., et al. (2016). Orphan crops browser: a bridge between model and orphan crops. *Mol. Breeding* 36, 9. doi: 10.1007/s11032-015-0430-2

Kans, J. (2013). *Entrez Direct: E-utilities on the Unix Command Line. National Center for Biotechnology Information (US)*. Available online at: https://www.ncbi.nlm.nih.gov/books/NBK179288/ (accessed March 20, 2023).

Kaur, P., Singh, A., and Chana, I. (2021). Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. *Arch Computat. Methods Eng.* 28, 4595–4631. doi: 10.1007/s11831-021-09547-0

Khayatbashi, S., Ferrada, S., and Hartig, O. (2022). "Converting property graphs to RDF: a preliminary study of the practical impact of different mappings," in *Proceedings of the 5th ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA) GRADES-NDA '22* (New York, NY: Association for Computing Machinery), 1–9. doi: 10.1145/3534540.3534695

Kreplak, J., Madoui, M. A., Cápal, P., Novák, P., and Labadie, K., Aubert, G., et al. (2019). A reference genome for pea provides insight into legume genome evolution. *Nat. Genet.* 51, 1411–1422. doi: 10.1038/s41588-019-0480-1

Krishnakumar, V., Kim, M., Rosen, B. D., Karamycheva, S., Bidwell, S. L., Tang, H., et al. (2015). MTGD: the medicago truncatula genome database. *Plant Cell Physiol.* 56, e1. doi: 10.1093/pcp/pcu179

Laporte, M. -A., Valette, L., Arnaud, E., Cooper, L., Meier, A., Jaiswal, P., et al. (2016). *Comparison of Ontology Mapping Techniques to Map Plant Trait Ontologies*. Corvallis, OR: CEUR Workshop Proceedings. Available online at: https://ceur-ws.org/Vol-1747/IP17_ICBO2016.pdf

Larmande, P., and Todorov, K. (2021). "AgroLD: A knowledge graph for the plant sciences," in *Semantic Web - ISWC 2021 Lecture Notes in Computer Science*, eds A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, and P. Barnaghi (Cham: Springer International Publishing), 496–510. doi: 10.1007/978-3-030-88361-4_29

Laurie, R. E., Diwadkar, P., Jaudal, M., Zhang, L., Hecht, V., Wen, J., et al. (2011). The Medicago flowering locus T Homolog, MtFTa1, Is a Key Regulator of Flowering *Time*. 156, 18. doi: 10.1104/pp.111.180182

Lees, J., Yeats, C., Perkins, J., Sillitoe, I., Rentzsch, R., Dessailly, B. H., et al. (2012). Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. *Nucleic Acids Res.* 40, D465–471. doi: 10.1093/nar/gkr1181

Li, J., Dai, X., Liu, T., and Zhao, P. X. (2012). LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.* 40, D1221–1229. doi: 10.1093/nar/gkr939

Li, J., Dai, X., Zhuang, Z., and Zhao, P. X. (2016). LegumeIP 2.0–a platform for the study of gene function and genome evolution in legumes. *Nucleic Acids Res.* 44, D1189–1194. doi: 10.1093/nar/gkv1237

Liao, Y., Wang, J., Zhu, Z., Liu, Y., Chen, J., Zhou, Y., et al. (2022). The 3D architecture of the pepper genome and its relationship to function and evolution. *Nat. Commun.* 13, 3479. doi: 10.1038/s41467-022-31112-x

Linard, B., Ebersberger, I., McGlynn, S. E., Glover, N., Mochizuki, T., Patricio, M., et al. (2021). Ten years of collaborative progress in the quest for orthologs. *Mol. Biol. Evol.* 38, 3033–3045. doi: 10.1093/molbev/msab098

Lohse, M., Nagel, A., Herter, T., May, P., Schroda, M., Zrenner, R., et al. (2014). Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data: Mercator: sequence functional annotation server. *Plant Cell Environ.* 37, 1250–1258. doi: 10.1111/pce.12231

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant J.* 53, 661–673. doi: 10.1111/j.1365-313X.2007.03326.x

Mi, H., and Thomas, P. (2009). PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol.* 563, 123–140. doi: 10.1007/978-1-60761-175-2_7

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi: 10.1530/JME-18-0055

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: the protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Naithani, S., Gupta, P., Preece, J., D'Eustachio, P., Elser, J. L., Garg, P., et al. (2019). Plant Reactome: a knowledgebase and resource for comparative pathway analysis. *Nucleic Acids Res.* 47, gkz996. doi: 10.1093/nar/gkz996

NCBI (2023). SRA Metadata and Submission Overview. Available online at: https://www.ncbi.nlm.nih.gov/sra/docs/submitmeta/ (accessed March 20, 2023).

Neo4j (2023a). Arrows. *Neo4j Graph Data Platform*. Available online at: https://neo4j.com/labs/arrows/ (accessed March 6, 2023).

Neo4j (2023b). *The Neo4j Graph Data Platform. Neo4j Graph Data Platform*. Available online at: https://neo4j.com/product/ (accessed January 30, 2023).

Oellrich, A., Walls, R. L., Cannon, E. K., Cannon, S. B., Cooper, L., Gardiner, J., et al. (2015). An ontology approach to comparative phenomics in plants. *Plant Methods* 11, 10. doi: 10.1186/s13007-015-0053-y

Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., et al. (2006). The rice annotation project database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic Acids Res.* 34, D741–D744. doi: 10.1093/nar/gkj094

Patel, H., Beber, M. E., Han, D. W., Philips, E., Manning, J., Yates, J. A. F., et al. (2022). nf-core/fetchngs: nf-core/fetchngs v1.9 - Plutonium Prancer. *Zenodo*. doi: 10.5281/zenodo.7468050

Patel, H., Ewels, P., Peltzer, A., Botvinnik, O., Sturm, G., Moreno, D., et al. (2023). nf-core/rnaseq: nf-core/rnaseq v3.10.1 – Plastered Rhodium Rudolph. *Zenodo*. doi: 10.5281/zenodo.7505987

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* 14, 417–419. doi: 10.1038/nmeth.4197

Pecrix, Y., Staton, S. E., Sallet, E., Lelandais-Brière, C., Moreau, S., Carrère, S., et al. (2018). Whole-genome landscape of Medicago truncatula symbiotic genes. *Nature Plants* 4, 1017–1025. doi: 10.1038/s41477-018-0286-7

Pham, G. M., Hamilton, J. P., Wood, J. C., Burke, J. T., Zhao, H., Vaillancourt, B., et al. (2020). Construction of a chromosome-scale long-read reference genome assembly for potato. *GigaScience* 9, giaa100. doi: 10.1093/gigascience/giaa100

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033

Raciti, D., Yook, K., Harris, T. W., Schedl, T., and Sternberg, P. W. (2018). *Micropublication* : incentivizing community curation and placing unpublished data into the public domain. *Database* 2018., e013 doi: 10.1093/database/bay013

Ramsay, L., Koh, C. S., Kagale, S., Gao, D., Kaur, S., Haile, T., et al. (2021). *Genomic rearrangements have consequences for introgression breeding as revealed by genome assemblies of wild and cultivated lentil species. Plant Biol.* 24, 237. doi: 10.1101/2021.07.23.453237

Roussey, C. (2021). *BBCH-based Plant Phenological Description Ontology*. doi: 10.15454/TIMQHW

Rubiales, D., Annicchiarico, P., Vaz Patto, M. C., and Julier, B. (2021). Legume breeding for the agroecological transition of global agri-food systems: a european perspective. *Front. Plant Sci.* 12, 782574. doi: 10.3389/fpls.2021.782574

Sanderson, L. A., Caron, C. T., Tan, R., Shen, Y., and Liu, R., Bett, K. E., et al. (2019). KnowPulse: a web-resource focused on diversity data for pulse crop improvement. *Front. Plant Sci.* 10, 965. doi: 10.3389/fpls.2019.00965

Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Strauss, M., Geyer, P. E., et al. (2022). *A knowledge graph to interpret clinical proteomics data. Nat Biotechnol.* 40, 692–702. doi: 10.1038/s41587-021-01145-6

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., et al. (2008). Genome structure of the legume, lotus japonicus. *DNA Res.* 15, 227–239. doi: 10.1093/dnares/dsn008

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26. doi: 10.1093/nar/gkab1112

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., et al. (2019). MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* 12, 879–892. doi: 10.1016/j.molp.2019.01.003

Semba, R. D., Ramsing, R., Rahman, N., Kraemer, K., and Bloem, M. W. (2021). Legumes as a sustainable source of protein in human diets. *Global Food Security* 28, 100520. doi: 10.1016/j.gfs.2021.100520

Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for fastA/Q file manipulation. *PLoS ONE* 11, e0163962. doi: 10.1371/journal.pone.0163962

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., et al. (2017). DNA sequencing at 40: past, present and future. *Nature* 550, 345–353. doi: 10.1038/nature24286

Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., et al. (2012). Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front. Physio.* 3, 326. doi: 10.3389/fphys.2012.00326

Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 51, D977–D985. doi: 10.1093/nar/gkac1010

Soneson, C., Love, M. I., and Robinson, M. D. (2016). *Differential analyses for* RNA-seq: transcript-level estimates improve gene-level inferences. doi: 10.12688/f1000research.7563.2

Stevens, I., Mukarram, A. K., Hörtenhuber, M., Meehan, T. F., Rung, J., Daub, C. O., et al. (2020). Ten simple rules for annotating sequencing experiments. *PLOS Computat. Biol.* 16, e1008260. doi: 10.1371/journal.pcbi.1008260

Su, X., Wang, B., Geng, X., Du, Y., Yang, Q., Liang, B., et al. (2021). A high-continuity and annotated tomato reference genome. *BMC Genomics* 22, 898. doi: 10.1186/s12864-021-08212-x

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., Paterson, A. H., et al. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tello-Ruiz, M. K., Naithani, S., Gupta, P., Olson, A., Wei, S., Preece, J., et al. (2021). Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.* 49, D1452–D1463. doi: 10.1093/nar/gkaa979

The Legume Phylogeny Working Group, Bruneau, A., Doyle, J. J., Herendeen, P., Hughes, C., Kenicer, G., et al. (2013). Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species–rich clades. *TAXON* 62, 217–248. doi: 10.12705/622.8

Thomson, G., Taylor, J., and Putterill, J. (2019). The transcriptomic response to a short day to long day shift in leaves of the reference legume *Medicago truncatula*. *PeerJ* 7, e6626. doi: 10.7717/peerj.6626

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y., et al. (2012). Dissecting plant genomes with the plaza comparative genomics platform. *Plant Physiol.* 158, 590–600. doi: 10.1104/pp.111.189514

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acid. Res.* 50, D1468–D1474. doi: 10.1093/nar/gkab1024

Venkatesan, A., Tagny Ngompe, G., Hassouni, N. E., Chentli, I., Guignon, V., Jonquet, C., et al. (2018). Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy. *PLoS ONE* 13, e0198270. doi: 10.1371/journal.pone.0198270

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010). "A comparison of a graph database and a relational database: a data provenance perspective," in *Proceedings of the 48th Annual Southeast Regional Conference on - ACM SE '10* (Oxford, MS: ACM Press). doi: 10.1145/1900008.1900067

Vizueta, J., Rozas, J., and Sánchez-Gracia, A. (2018). Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biol. Evol.* 10, 1221–1236. doi: 10.1093/gbe/evy081

W3C (1994). Available online at: https://www.w3.org/ (accessed February 15, 2023).

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, e49. doi: 10.1093/nar/gkr1293

Wei, Q., Wang, J., Wang, W., Hu, T., Hu, H., Bao, C., et al. (2020). A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic Res.* 7, 153. doi: 10.1038/s41438-020-00391-0

Wilkinson, M. D., and Dumontier, M., Aalbersberg, I.j., J., Appleton, G., Axton, M., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018. doi: 10.1038/sdata.2016.18

Williams, O., Vander Schoor, J. K., Butler, J. B., Ridge, S., Sussmilch, F. C., Hecht, V. F. G., et al. (2022). The genetic architecture of flowering time changes in pea from wild to crop. *J. Exp. Bot.* 73, 3978–3990. doi: 10.1093/jxb/erac132

Xi, H., Nguyen, V., Ward, C., Liu, Z., and Searle, I. R. (2022). Chromosome-level assembly of the common vetch (Vicia sativa) reference genome. *Gigabyte* 2022, 1–20. doi: 10.46471/gigabyte.38

Yang, T., Liu, R., Luo, Y., Hu, S., Wang, D., Wang, C., et al. (2022). *Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics*. *Nat Genet.* 10, 1553–1563. doi: 10.1038/s41588-022-01172-2

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., et al. (2022). Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* 50, D996–D1003. doi: 10.1093/nar/gkab1007

Ye, C. -Y., and Fan, L. (2021). Orphan crops and their wild relatives in the genomic era. *Mol. Plant* 14, 27–39. doi: 10.1016/j.molp.2020.12.013

Yoon, B. H., Kim, S-, K., and Kim, S. Y. (2017). Use of graph database for the integration of heterogeneous biological data. *Genomics Inform* 15, 19. doi: 10.5808/GI.2017.15.1.19

Yuan, H. Y., Caron, C. T., Ramsay, L., and Fratini, R., de la Vega, M. P., Vandenberg, A., et al. (2021). Genetic and gene expression analysis of flowering time regulation by light quality in lentil. *Annal. Bot.* 128, 481–496. doi: 10.1093/aob/mcab083

TYPE Methods
PUBLISHED 27 September 2023
DOI 10.3389/frai.2023.1188036

OPEN ACCESS

EDITED BY
Qichun Yang,
The University of Melbourne, Australia

REVIEWED BY
Shuai Xie,
Changjiang River Scientific Research Institute (CRSRI), China
Angela Kranz,
Helmholtz Association of German Research Centres (HZ), Germany

*CORRESPONDENCE
Catherine Roussey
✉ catherine.roussey@inrae.fr

RECEIVED 16 March 2023
ACCEPTED 14 August 2023
PUBLISHED 27 September 2023

CITATION
Amardeilh F, Aubin S, Bernard S, Bravo S, Bossy R, Faron C, Michel F, Raphel J and Roussey C (2023) Combining different points of view on plant descriptions: mapping agricultural plant roles and biological taxa. *Front. Artif. Intell.* 6:1188036.
doi: 10.3389/frai.2023.1188036

Combining different points of view on plant descriptions: mapping agricultural plant roles and biological taxa

Florence Amardeilh[1], Sophie Aubin[2], Stephan Bernard[3], Sonia Bravo[2], Robert Bossy[4], Catherine Faron[5], Franck Michel[5], Juliette Raphel[1] and Catherine Roussey[3,6]*

[1]Elzeard, Bordeaux, France, [2]DipSO, INRAE, Paris, France, [3]Université Clermont Auvergne, INRAE, UR TSCF, Clermont-Ferrand, France, [4]MaIAGE, INRAE, Université Paris-Saclay, Jouy-en-Josas, France, [5]Université Côte d'Azur, Inria, I3S, Sophia-Antipolis, France, [6]MISTEA, University of Montpellier, INRAE & Institut Agro, Montpellier, France

This article describes our study on the alignment of two complementary knowledge graphs useful in agriculture: the thesaurus of cultivated plants in France named French Crop Usage (FCU) and the French national taxonomic repository TAXREF for fauna, flora, and fungi. FCU describes the usages of plants in agriculture: "*tomatoes*" are crops used for human food, and "*grapevines*" are crops used for human beverage. TAXREF describes biological taxa and associated scientific names: for example, a tomato species may be "*Solanum lycopersicum*" or a grapevine species may be "*Vitis vinifera*". Both knowledge graphs contain vernacular names of plants but those names are ambiguous. Thus, a group of agricultural experts produced some mappings from FCU crops to TAXREF taxa. Moreover, new RDF properties have been defined to declare those new types of mapping relations between plant descriptions. The metadata for the mappings and the mapping set are encoded with the Simple Standard for Sharing Ontological Mappings (SSSOM), a new model which, among other qualities, offers means to report on provenance of particular interest for this study. The produced mappings are available for download in Recherche Data Gouv, the federated national platform for research data in France.

KEYWORDS

mapping, SSSOM, crop, taxa, thesaurus French Crop Usage, French national taxonomic repository, TAXREF, SKOS model

1. Introduction

While the Web of linked data makes more and more knowledge graphs available, their cross-use often remains a challenge. This study presents a dataset containing mappings between two knowledge graphs representing different points of view on the same objects. This mapping set should allow to query simultaneously these graphs to enrich object descriptions by combining these points of view. Agriculture offers a particular use case of mappings, linked to the modeling of cultivated plants. Several expertises are needed to describe a cultivated plant: farmer vs. agronomist, agronomist vs. ecologist. The scientific world (ecologists or agronomists) tends to use scientific names from taxonomic science to designate living organisms (plants, insects). These scientific names are stored in biological taxonomies. The world of users (farmers) generally uses vernacular names or domain specific categories (e.g., cereals) to designate the living organisms involved in their practice. In parallel, a plant can have several usages in agriculture: (1) plants cultivated in a plot for production purposes such as vegetables or cereals, in other word crops, (2) weeds that appear on a plot without being cultivated for which farmers want to limit the development or remove them from the plot, (3) a first cultivated plant that provides a service to a second cultivated plant for production purposes. Both plants are cultivated on the same plot but sometime not at the same time. The first cultivated plant will be destroyed without being harvested and is called service plant. The second cultivated plant will be harvested and is called the crop. For reasons of conciseness, we will limit this article to the plant usages for production purposes that is to say crops.

We present our study on the mappings of two complementary knowledge graphs useful in the agricultural domain: the French Crop Usage thesaurus (FCU) and the French national taxonomic register TAXREF for fauna, flora, and fungi. FCU describes the usage of plants in agriculture: "*tomatoes*" are crops used for human food, "*grapevines*" are crops used for human food or beverage. It represents the farmers' point of view. TAXREF describes biological taxa and associated scientific names: for example, a tomato species may be "*Solanum lycopersicum*" or a grapevine species may be "*Vitis vinifera*". TAXREF represents the agronomists' point of view. Both knowledge graphs contain vernacular names of plants. Vernacular names are often ambiguous and not consensual, which renders the matching activity particularly challenging.

Our previous studies (Michel et al., 2022) have implemented several automatic alignment methods based on vernacular names comparison. Those automatic methods reused existing reference sources such as EPPO global database[1] and the official French catalog of species and varieties of cultivated plants GEVES.[2] The results show that it is necessary to clean the automatically produced alignments due to the ambiguity of vernacular names. Therefore, a group of agricultural experts has produced a set of valid mappings. Those mappings are published as open data on the French Recherche Data Gouv repository.[3] Thus, they could be used as a gold standard to validate any automatic alignment methods.

The remainder of the study is organized as follows: Section 2 describes first the knowledge graphs and vocabularies used in our mapping set (Section 2.1), followed by the manual method applied to align the two knowledge graphs (Section 2.2). Section 3 presents our analysis of the challenge encountered in matching the graphs and representing the mapping set using the SSSOM model. In Section 4, we summarize the results and provide an outlook for future improvement.

2. Methods

2.1. Materials

First, we describe in detail the two aligned knowledge graphs: TAXREF-LD and FCU. Second we present the RDF vocabulary that we defined to declare new types of mapping relations between plant descriptions. Indeed, SKOS properties are not sufficient to align an agricultural usage with a scientific taxon. Third, the SSSOM vocabulary is presented to store the set of mappings and their metadata.

2.1.1. TAXREF and TAXREF-LD

TAXREF (Gargominy et al., 2021) is the French taxonomic repository for fauna, flora, and fungi. In addition to a Web portal, a REST service, and a set of downloadable CSV files, TAXREF is available in the form of a knowledge graph complying with the Linked Data principles, named TAXREF-LD (Michel et al., 2017). TAXREF-LD is available on the AgroPortal repository.[4] This study has been developed using the 15.2 version of TAXREF-LD which contains 287,229 classes and more than 1,000,000 instances.

To accurately reflect the distinction between taxonomy (a taxon gathers biological individuals that share common characteristics) and nomenclature (the scientific names assigned to taxa), TAXREF-LD has two distinct levels of modeling, as shown in Figure 1. At the taxonomic level, each taxon is modeled as an OWL class whose members are the biological individuals of that taxon. The parent class is the higher ranked taxon (e.g., "*Daucus carota*" is of rank species, the parent class "*Daucus*" is of rank genus). At the nomenclatural level, scientific names are represented as concepts in a SKOS thesaurus. Each name (instance of *skos:Concept*) is linked to a taxon (an *OWL class*) by a property indicating whether it is the reference name (*accepted name* in zoology or *valid name* in botany) or a synonym. The figure also presents vernacular names that are represented as a simple literal as well as a blank node of type *skos-xl:Label* that reifies the vernacular name and makes it possible to provide additional information such as the geographic area in which this vernacular name is valid or a bibliographic reference. In addition to strictly taxonomic information, TAXREF-LD also represents other types of information not shown on this figure, such as habitats, conservation status, biogeographical status, interactions between species, and the bibliographical references associated with this information. Notably, TAXREF-LD sometimes associates the same vernacular name with several taxa. These vernacular names are taken from the publications where the scientific names are declared. Furthermore, TAXREF-LD is linked to several third-party taxonomic repositories including Agrovoc Thesaurus and NCBI Organismal Taxonomy.

2.1.2. French Crop Usage thesaurus

The French Crop Usage (FCU) thesaurus normalizes crop names in French. Moreover, it organizes these crop names in categories, according to their usages on the French territory. The usages represent also the agricultural sectors.

As shown in Figure 2, the thesaurus hierarchy has two main branches. The branch named "*Multiusages*" contains all the cultivated plants that have several usages in agriculture. For example, "*carotte*" (carrot) may be used as vegetable or fodder. The branch "*Usages_plantes_cultivees*" organizes cultivated plants according to their usages and represents agricultural sectors. In this

[1] https://gd.eppo.int/

[2] Catalogue officiel des espèces et variétés de plantes cultivées en France accessible at https://www.geves.fr/catalogue/.

[3] https://doi.org/10.57745/LVRFWJ

[4] https://agroportal.lirmm.fr/ontologies/TAXREF-LD

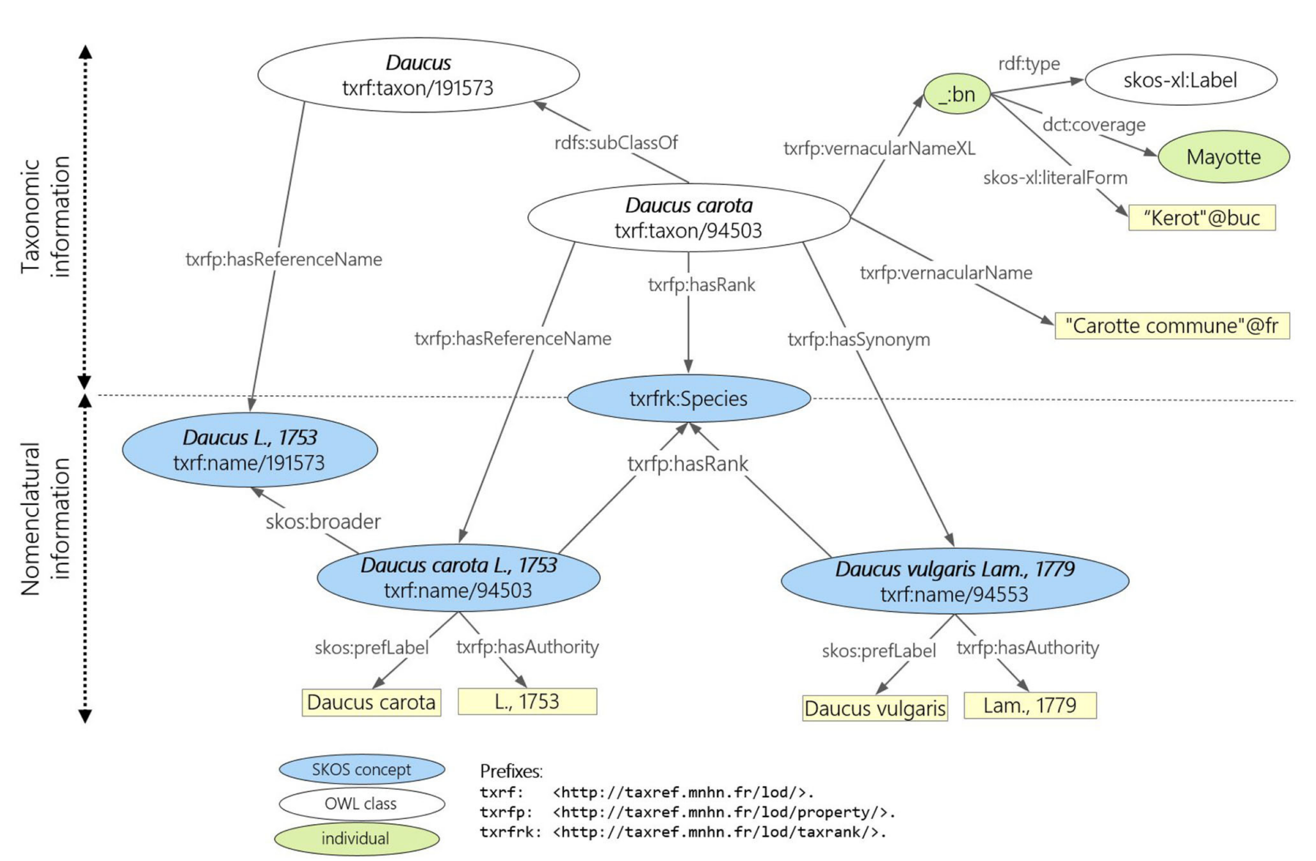


FIGURE 1
Modeling principle of TAXREF-LD distinguishing taxonomic and nomenclatural information. The species taxon "*Daucus carota*" has for scientific name "*Daucus carota* L, 1758".

branch, the crop usage "*carotte potagère*" is linked to the vegetable category "*légume racine*" (root vegetable).

The FCU thesaurus is formalized using the Simple Knowledge Organization System (SKOS) vocabulary proposed by W3C (Miles and Bechhofer, 2009). Each crop usage or category is represented by an instance of *skos:Concept*. The thesaurus is published on the Web using Linked Data principles.

The thesaurus is available on the AgroPortal repository.[5] This study has been developed using the 3.3 version of FCU which contains 707 instances of *skos:Concept*. The maximum depth of the hierarchy is 6 levels. Each *skos:Concept* is defined by several properties, as shown in Figure 3. The description of a crop usage or category contains the following:

- The value of property *skos:prefLabel* is the crop name in French. The term is the vernacular name of the cultivated plant or the category name. To avoid ambiguity in the case of a cultivated plant with different usages, the crop name is the combination of the vernacular name of the plant and its usage. For example, in Figure 3, the crop name is "*carotte*" + "*potagère*".
- The value of property *skos:altLabel* is other possible labels that can be used for the crop. For example, in Figure 3, an alternative crop name is "*carotte cultivée*".
- The value of property *skos:definition* is the definition of the crop usage in French. The definition accounts for the crop position in the hierarchy.
- The value of property *skos:note* is at least one definition from another source, such as the French Wikipedia. The definition always ends by the indication of the source. For example, in Figure 3, the crop "*carotte potagère*" was found in The *Official Catalog of Species and Varieties of Cultivated Crops in France*.[6] Thus, depending on the source, the same crop may have different names which show the ambiguity of crop names.

2.1.3. Mapping properties: taxon vs. usage

In TAXREF-LD, a taxon is defined by an *OWL class*, and the names of taxon are defined by instances of *skos:Concept*. In FCU, a crop usage is defined by an instance of *skos:Concept*.

We have defined 10 annotation properties to link an *OWL class* representing a taxon to an instance of *skos:Concept* representing a crop usage. The main annotation property is *ontofcu:hasTaxon* (/ its inverse property is *ontofcu:hasUsage*). This property (/ its inverse property) links a crop usage to a taxon. This relation indicates that the taxon is a candidate to fulfill the crop usage. For example, the species "*Daucus Carota*" can be used as "*carotte fourragère*" (fodder

[5] https://agroportal.lirmm.fr/ontologies/CROPUSAGE/?p=summary

[6] https://www.geves.fr/catalogue-france/

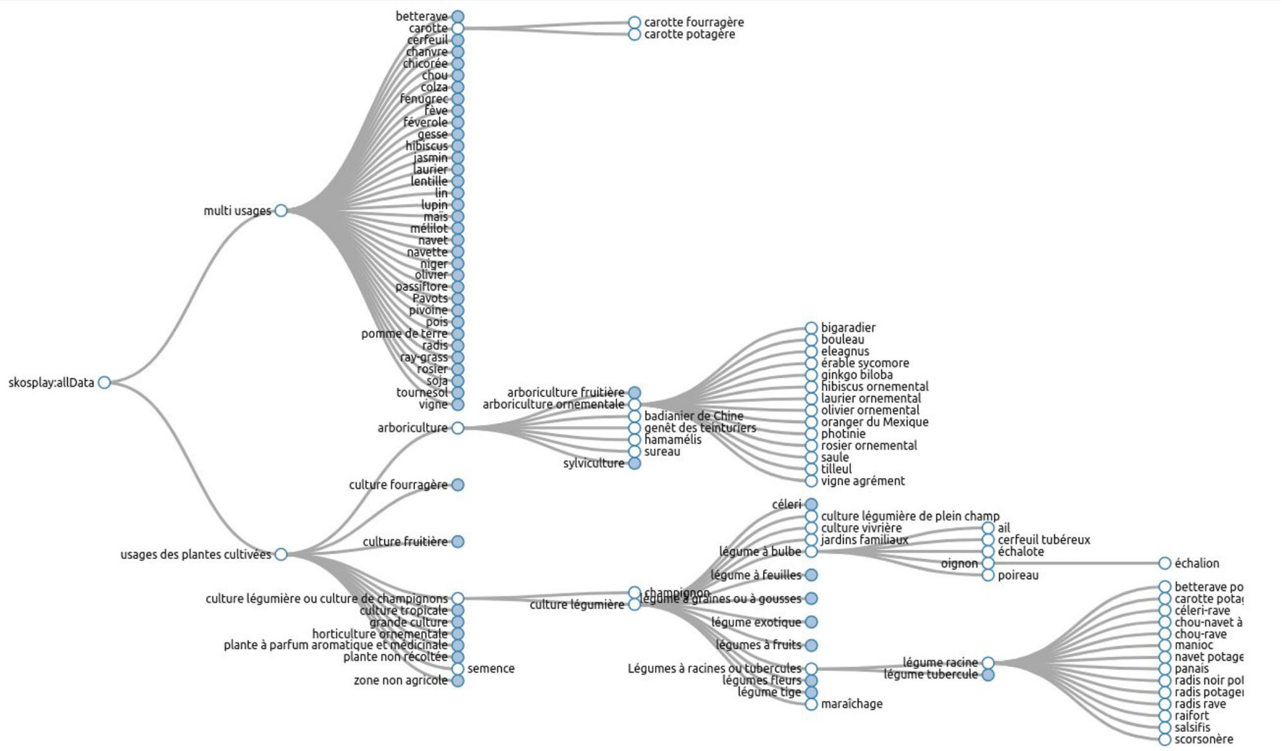
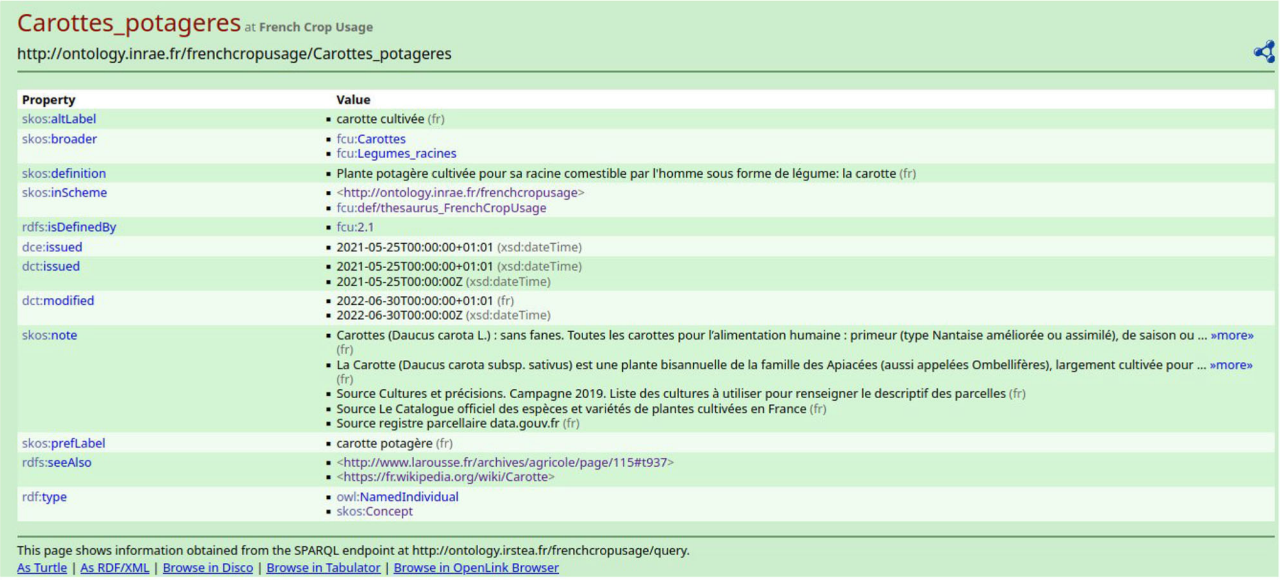


FIGURE 2
An extract from the FCU thesaurus, visualized with the SKOS Play tool.

carrot). This property is specialized into four annotation properties (and their inverse) as follows:

- *ontofcu:hasGenericTaxon* (/ *ontofcu:hasSpecificUsage*) annotation property represents a relationship from a crop usage to a taxon. This relation indicates that one of the descendants of the taxon is the reference taxon to fulfill the crop usage. It is used when the descendant is not defined in the taxonomy source (e.g., TAXREF-LD). For example, the form "*Cichorium intybus var. foliosum Hegi f. cylindricum*" is known to fulfill the vegetable usage "*chicorée pain de sucre*" (sugarloaf chicory). Unfortunately, this form does not appear in TAXREF-LD. Its parent, the variety "*Cichorium intybus var. Foliosum*," belongs to TAXREF-LD. Thus, the mapping between "*chicorée pain de sucre*" (sugarloaf chicory) and the variety "*Cichorium intybus var. Foliosum*" will use the property *ontofcu:hasGenericTaxon*.
- *ontofcu:hasInvalidTaxon* (/ *ontofcu:hasInvalidUsage*) annotation property represents a relationship from a crop usage to a taxon. This relation indicates that the taxon can not be used to fulfill the crop usage. For example, the subspecies "*Daucus carota subsp. gadecaei*" is not a cultivated plant and can not be used as "*carotte potagère*" (vegetable carrot). This property is used to invalidate the output of automatic alignment tool.
- *ontofcu:hasReferenceTaxon* (/ *ontofcu:hasReferenceUsage*) annotation property represents a relationship from a crop usage to a taxon. This relation indicates that the taxon is the reference known to fulfill the crop usage. For example, the species "*Daucus Carota*" is one of the reference taxa used as "*carotte potagère*" (vegetable carrot). The subspecies "*Daucus carota subsp. sativus*" is another reference taxon to be used as "*carotte potagère*" (vegetable carrot).
- *ontofcu:hasSpecificTaxon* (/ *ontofcu:hasGenericUsage*) annotation property represents a relationship from a crop usage to a taxon. This relation indicates that one of the descendants of the crop usage is the reference usage of the taxon. It is used when the descendant of the crop usage is not defined in FCU. For example, the variety "*Solanum lycopersicum var. cerasiforme*" is known to fulfill the crop usage "*tomate cerise*" (cherry tomato). Unfortunately, this type of tomato is not defined in FCU. Thus, the mapping between "*tomate*" (tomato) and the variety "*Solanum lycopersicum var. cerasiforme*" will use the property *ontofcu:hasSpecificTaxon*.

We have defined 12 object properties to link an instance of *skos:Concept* representing a scientific name to an instance of *skos:Concept* representing a crop usage. Those object property triples should be associated with annotation property triples listed above used as documentation. The main object property is *ontofcu:hasScientificName* (/ *ontofcu:hasVernacularName*). This property (/ its inverse property) is a relation from a crop usage to a taxon scientific name. Both are represented as an instance of *skos:Concept*. This relation indicates that the taxon scientific name is a candidate to identify the crop usage. For example, "*Daucus carota L., 1753*" can be the scientific name of the crop usage



Carottes_potageres at French Crop Usage

http://ontology.inrae.fr/frenchcropusage/Carottes_potageres

Property	Value
skos:altLabel	▪ carotte cultivée (fr)
skos:broader	▪ fcu:Carottes ▪ fcu:Legumes_racines
skos:definition	▪ Plante potagère cultivée pour sa racine comestible par l'homme sous forme de légume: la carotte (fr)
skos:inScheme	▪ <http://ontology.inrae.fr/frenchcropusage> ▪ fcu:def/thesaurus_FrenchCropUsage
rdfs:isDefinedBy	▪ fcu:2.1
dce:issued	▪ 2021-05-25T00:00:00+01:01 (xsd:dateTime)
dct:issued	▪ 2021-05-25T00:00:00+01:01 (xsd:dateTime) ▪ 2021-05-25T00:00:00Z (xsd:dateTime)
dct:modified	▪ 2022-06-30T00:00:00+01:01 (fr) ▪ 2022-06-30T00:00:00Z (xsd:dateTime)
skos:note	▪ Carottes (Daucus carota L.) : sans fanes. Toutes les carottes pour l'alimentation humaine : primeur (type Nantaise améliorée ou assimilé), de saison ou ... »more» (fr) ▪ La Carotte (Daucus carota subsp. sativus) est une plante bisannuelle de la famille des Apiacées (aussi appelées Ombellifères), largement cultivée pour ... »more» (fr) ▪ Source Cultures et précisions. Campagne 2019. Liste des cultures à utiliser pour renseigner le descriptif des parcelles (fr) ▪ Source Le Catalogue officiel des espèces et variétés de plantes cultivées en France (fr) ▪ Source registre parcellaire data.gouv.fr (fr)
skos:prefLabel	▪ carotte potagère (fr)
rdfs:seeAlso	▪ <http://www.larousse.fr/archives/agricole/page/115#t937> ▪ <https://fr.wikipedia.org/wiki/Carotte>
rdf:type	▪ owl:NamedIndividual ▪ skos:Concept

This page shows information obtained from the SPARQL endpoint at http://ontology.irstea.fr/frenchcropusage/query.

As Turtle | As RDF/XML | Browse in Disco | Browse in Tabulator | Browse in OpenLink Browser

FIGURE 3
The information related to the *skos:Concept* instance "*fcu:Carottes_potageres*".

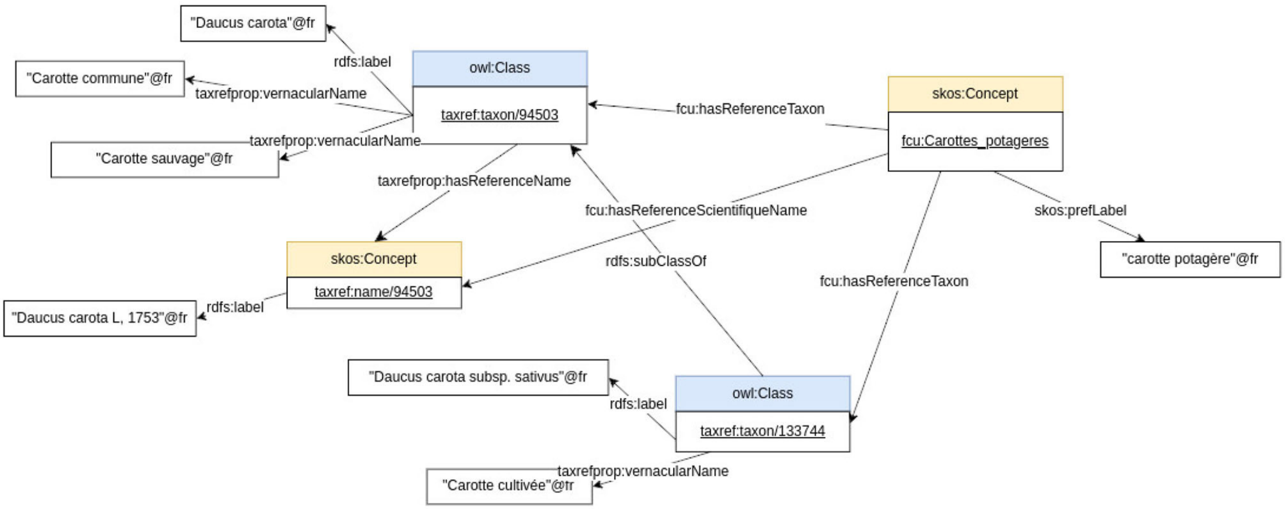


FIGURE 4
Some mappings between the crop usage "*carotte potagère*", the taxon "*Daucus carota*", its scientific name "*Daucus carota* L, 1758", and the taxon "*Daucus carota* subsp. *sativus*" using the CHOWLK language.

"*carotte fourragère*" (fodder carrot). Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasTaxon / ontofcu:hasUsage*.

- *ontofcu:hasGenericScientificName* (/ *ontofcu:hasSpecificVernacularName*) object property represents a relationship from a crop usage to a taxon scientific name. This relation indicates that one of the descendants of the taxon name is known to be the scientific name of the crop usage. This is used when the descendant is not defined in the taxonomy. For example, the form scientific name "*Cichorium intybus var. foliosum Hegi f. cylindricum*" is known to be the scientific name of the vegetable usage "*chicorée pain de sucre*" (sugarloaf chicory). Unfortunately, this form does not belong to TAXREF-LD. But its parent, the variety scientific name "*Cichorium intybus var. Foliosum Hegi, 1928*" belongs to TAXREF-LD. Thus, the mapping between "*chicorée pain de sucre*" (sugarloaf chicory) and the variety scientific name "*Cichorium intybus var. Foliosum Hegi, 1928*" will use the property *ontofcu:hasGenericScientificName*. Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasGenericTaxon*.
- *ontofcu:hasInvalidScientificName* (/ *hasInvalidVernacularName*) object property represents a relationship from a crop usage to a taxon scientific name. This relation indicates that the taxon scientific name can not be

used to identify the crop usage. For example, the subspecies "*Daucus carota subsp. gadecaei*" is not a cultivated plant. Thus, "*Daucus carota subsp. gadecaei (Rouy & E.G.Camus) Heywood, 1968*" is not the scientific name of "*carotte potagère*" (vegetable carrot). Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasInvalidTaxon*.

- *ontofcu:hasReferenceScientificName* (/ *ontofcu:hasReferenceVernacularName*) object property represents a relationship from a crop usage to a taxon scientific name. This relation indicates that the taxon scientific name can be used to identify the crop usage. For example, "*Daucus carota L., 1753*" is the reference scientific name of "*carotte fourragère*" (fodder carrot). Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasReferenceTaxon*.
- *ontofcu:hasSpecificScientificName* (/ *ontofcu:hasGenericVernacularName*) object property represents a relationship from a crop usage to a taxon scientific name. This relation indicates that one of the descendants of the crop usage is the reference vernacular name of the taxon. This is used when the descendant of the crop usage is not defined in FCU. For example, "*Solanum lycopersicum var. cerasiforme (Alef.) Fosberg, 1955*" is known to be the scientific name of the crop usage "*tomate cerise*" (cherry tomato). Unfortunately, "*tomate cerise*" does not belong to FCU. Thus, the mapping between "*tomate*" (tomato) and the variety scientific name "*Solanum lycopersicum var. cerasiforme (Alef.) Fosberg, 1955*" will use the property *ontofcu:hasSpecificScientificName*. Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasSpecificTaxon*.
- *ontofcu:hasSynonymousScientificName* object property represents a relationship from a crop usage to a taxon scientific name. This relation indicates that the scientific name is not the reference name but one of its synonyms and can also be used to identify the crop usage. For example, "*Daucus communis Rouy & E.G.Camus, 1901*" is the synonymous scientific name of "*carotte fourragère*" (fodder carrot). Notably, the crop usage and the associated taxon should also be linked by the annotation property *ontofcu:hasReferenceTaxon*.
- *ontofcu:hasSynonymousVernacularName* object property represents a relationship from a taxon scientific name to a crop usage. This relation indicates that the crop usage is the synonymous vernacular name of the taxon scientific name. For example, FCU contains a new collection of crop usage dedicated to subsistence crops. A new crop usage "*carotte du jardin*" (garden carrot) is added which is defined as synonymous (same as) to "*carotte potagère*" (vegetable carrot). Thus, "*carotte du jardin*" (garden carrot) will be the synonymous vernacular name of "*Daucus carota L., 1753*." Notably, the both crop usages and the associated taxon should also be linked by the annotation property *ontofcu:hasReferenceTaxon*.

Figure 4 shows an exception of the mapping between the FCU concept "*carotte potagère*" and two TAXREF-LD classes "*Daucus carota*" and "*Daucus carota subsp. sativus.*"[7]

Those properties are visible on the property tab of AgroPortal.[8]

The generic properties *ontofcu:hasTaxon* *ontofcu:hasScientificName* (/ *ontofcu:hasUsage* *ontofcu:hasVernacularName*) can be used to declare any automatic mappings, potentially associated with their score but without validation concern. The generic properties just state that mappings were produced by any automatic method, and they are candidate mappings that need validation. The specific properties will be used to declare valid, cleaned, and precised mappings with their confidence value.

2.1.4. Alignment metadata model SSSOM

Simple standard for Sharing Ontology Mappings (SSSOM) is a recent standard model developed by the biomedical community around OBO Foundry and described by Matentzoglu et al. (2022). It provides a rich set of metadata to describe mappings (individual mappings between a pair of entities) and a mapping set (a set of individual mappings). For this work, we used SSSOM version 0.15 that was released in July 2023.

The main objective of the SSSOM project is to propose a catalog of metadata allowing to have information about the provenance of the mappings, whether they are calculated manually or automatically. The expected impact is an augmented trustworthiness resulting in an increased reuse of mappings by third parties. Meanwhile, there is also a real desire to produce a model that is easy to use. Indeed, the project proposes, on the one hand, the serialization in RDF/OWL for the Semantic Web community and, on the other hand, a TSV format for a larger community which can thus exchange mappings in a simple format yet with rich semantics.

We chose to publish the mapping set in the TSV format as a first step. We opted for the embedded mode where mapping set level metadata are integrated in the mapping TSV file as commented YAML (prefixed with #). The properties used to describe the mapping set are presented in Table 1 while those used to describe each mapping are shown in Table 2. Mandatory properties are marked with an asterisk (*). For each property, we indicate its description as stated in the SSSOM model version 0.15. In Section 3, we present a short analysis of advantages and limits of the SSSOM model.

2.2. Manual alignment method

Previously, we have tested some automatic alignment methods (Michel et al., 2022) that reused existing reference sources such as EPPO global database (see text footnote[1]) and the official French catalog of species and varieties of cultivated plants GEVES (see text footnote[2]). The final automatic method computes a confidence score of the mapping between a taxon from TAXREF-LD and a

[7] The UML based CHOWLK language is used to present the knowledge graph. More information available on https://chowlk.linkeddata.es/notation.html.

[8] https://agroportal.lirmm.fr/ontologies/CROPUSAGE/?p=properties

TABLE 1 SSSOM (V0.15) properties used to describe our mapping set.

Property	Description	Example
mapping_set_id	A globally unique identifier for the mapping set (not each individual mapping). Should be IRI, ideally resolvable.	"https://doi.org/10.57745/LVRFWJ"
creator_id	Identifies the persons or groups responsible for the creation of the mapping. The creator is the agent that put the mapping in its published form, which may be different from the author, which is a person that was actively involved in the assertion of the mapping.	#creator_id: "https://ror.org/01pd2sz18"
creator_label	A string identifying the creator of this mapping.	#creator_label: "Mathématiques, Informatique et Statistique pour l'Environnement et l'Agronomie"
curie_map	A valid curie map that allows the unambiguous interpretation of CURIEs.	#curie_map : #fcu: "http://ontology.inrae.fr/frenchcropusage" #taxref: "http://taxref.mnhn.fr/lod/taxref-ld"
subject_source	URI of ontology source for the subject.	#subject_source: "http://ontology.inrae.fr/frenchcropusage"
subject_source_version	Version IRI or version string of the source of the subject term.	#subject_source_version: "3.3"
object_source	IRI of ontology source for the object. Version IRI preferred.	#object_source: "http://taxref.mnhn.fr/lod/taxref-ld"
object_source_version	Version IRI or version string of the source of the object term.	#object_source_version: "15.2"
license	A url to the license of the mapping. In absence of a license we assume no license.	#license: "https://creativecommons.org/licenses/by/2.0/"

crop usage from FCU using EPPO database and GEVES catalog. The main problems come from name ambiguity: depending on the source (1) the scientific name may follow or not the botanical nomenclature code used in TAXREF-LD. For example, the scientific name of carrot may be "*Daucus carota*", "*Daucus carota* L", or "*Daucus carota* L, 1753". (2) The vernacular name presents in the source may not match exactly the vernacular name displayed in FCU. For example, the vernacular name of carrot presents in the source may be "carotte sauvage" ("wild carrot"). This name does not appear in FCU. (3) The vernacular name present in the source may not identify precisely a crop usage, that is to say a FCU concept narrower than "*Usages_plantes_cultivees*". For example, carrot has two crop usages in FCU "carotte potagère" ("vegetable carrot") and "carotte fourragère" ("folder carrot"). The output of the automatic method should be cleaned by human experts due to the inaccuracy of name comparison.

Thus, we decided to create a new mapping set by asking experts to propose correct and well-known mappings between crop usages, taxa, and their scientific names. All the proposed mappings should have a high confidence value. If any ambiguity existed, the mapping should not be created. First, we provided the experts with guidelines to help them in their decisions. Second, some research tools were proposed to search terms into the two knowledge graphs. Third, three curation rules were written to contextualize the mappings they created and indicate the provenance of the mappings. We focused on specific crops: grapevine, carrot, chicory, and tomato according to the availability of experts.

As shown in Figure 5, two kind of experts are involved. First, the mapping reviewer proposes some mapping specifications based on its knowledge or other information source. A mapping specification looks like : the taxon "*Daucus carota*" is used as "carotte potagère" ("vegetable carrot"). Second, the reviewer author has to find the correct URI from TAXREF-LD and FCU knowledge graphs to produce the SSSOM mappings following the mapping specifications.

2.2.1. Generic guidelines

We provide the following guidelines to help the experts create their mappings:

1. Only the mappings from FCU crop usages to TAXREF-LD taxa are represented. We focus on crop usages that belong to the branch "*Usages_plantes_cultivees*" to avoid ambiguity. The goal is to select the most specific crop usage from FCU and align it to some TAXREF-LD taxa using the properties defined in Section 2.1.3.
2. If possible the experts should select a reference information source used to identify the mappings. Each information source is associated with a curation rule to describe the mapping identification method.
3. First, the experts should create a mapping between a crop usage and its reference taxon using the annotation property *ontofcu:hasReferenceTaxon*. We impose that the first reference taxon has the rank species.
4. Second, if a more specific type of taxon, for example, a subspecies or a variety, is well known as the reference taxon of the crop usage, another mapping is created using the annotation property *ontofcu:hasReferenceTaxon*.
5. If the known reference taxon of the crop usage is not available in TAXREF-LD, another type of mapping should be used. A triple based on property *ontofcu:hasGenericTaxon* should be created.
6. Based on the above annotation properties, the mapping between the crop usage and the taxon scientific names

TABLE 2 SSSOM (V0.15) properties used to describe our mappings.

Property	Description	Example
subject_id*	The ID of the subject of the mapping.	fcu:Carottes_fourrageres
subject_label	The label of subject of the mapping.	carotte fourragère
predicate_id*	The ID of the predicate or relation that relates the subject and object of this match.	fcu:def/hasReferenceTaxon
predicate_label	The label of the predicate/relation of the mapping.	Has reference taxon
object_id*	The ID of the object of the mapping.	taxref:taxon/133744
object_label	The label of object of the mapping.	*Daucus carota* subsp.*sativus*
confidence	A score between 0 and 1 to denote the confidence or probability that the match is correct, where 1 denotes total confidence.	1
mapping_justification	A mapping justification is an action (or the written representation of that action) of showing a mapping to be right or reasonable.	semapv: ManualMappingCuration
mapping_cardinality	A string indicating whether this mapping is from a 1:1 (the subject_id maps to a single object_id), 1:n (the subject maps to more than one object_id), n:1, 1:0, 0:1 or n:n group. Note that this is a convenience field that should be derivable from the mapping set.	N:N
subject_type	The type of entity that is being mapped.	skos:Concept
object_type	The type of entity that is being mapped.	owl:Class
author_id	Identifies the persons or groups responsible for asserting the mappings. Recommended to be a (pipe-separated) list of ORCIDs or otherwise identifying URLs, but any identifying string (such as name and affiliation) is permissible.	https://orcid.org/0000-0002-3076-5499
author_label	A string identifying the author of this mapping. In the spirit of provenance, consider to use author_id instead.	Catherine Roussey
reviewer_id	Identifies the persons or groups that reviewed and confirmed the mapping. Recommended to be a (pipe-separated) list of ORCIDs or otherwise identifying URLs, but any identifying string (such as name and affiliation) is permissible.	https://orcid.org/0000-0002-5872-5034
reviewer_label	A string identifying the reviewer of this mapping. In the spirit of provenance, consider to use author_id instead.	Juliette Raphel
mapping_date	The date the mapping was asserted. This is different from the date the mapping was published or compiled in a SSSOM file.	2023-02-03T00:00:00Z
curation_rule_text	The textual representation of curation rule is intended to be used in cases where (1) the creation of a resource is not practical from the perspective of the mapping_provider and (2) as an additional piece of metadata to augment the curation_rule element with a human readable text.	CR_Experts: the experts found the mapping in the Official Catalog of Species and Varieties of Cultivated Crops in France [https://www.geves.fr/catalogue-france/]
comment	Free text field containing either curator notes or text generated by tool providing additional informative information.	Many subspecies of *Daucus carota* can be eaten and used as forage : the most well known is *Daucus carota* subsp. *sativus*

are derived. First, a mapping is created between the crop usage and the reference scientific name using the object property *ontofcu:hasReferenceScientificName*. In TAXREF-LD, the reference scientific name of a taxon can be found by following the link *taxref:hasReferenceName*.

7. If the reference source of information (or the experts' knowledge) indicates another scientific name than the one provided in TAXREF-LD as reference name, the experts should try to find it in the synonym names of the taxon and create a new mapping between the crop usage and the scientific name using the object property *ontofcu:hasSynonymeScientificName*.

The name of the experts are indicated in mapping information as mapping reviewer or mapping author. The reviewer is the expert who searched into the reference source the mapping information (or who knew for sure the taxon that fulfills the crop usage). The author is the expert who searched into the knowledge graph the URI entity based on reviewer information. The reviewer information is synthesized into the comment of the mapping (see Table 2).

2.2.2. Search tools

To find a crop usage in FCU thesaurus, the following query interfaces are available:

One simple solution is to navigate through the AgroPortal interface concept tab,[9] to select the most specific instance of *skos:Concept*. Figure 6 shows the hierarchy exploration to find the

[9] https://agroportal.lirmm.fr/ontologies/CROPUSAGE/?p=classes

instance named "*fcu:Carottes_potageres*." Notably, this interface presents first English labels, and if no English label is provided, the French labels are presented. Thus, Figure 6 presents a mixture of English and French labels.

To be sure to search French labels, another solution is to query the SPARQL Endpoint.[10] For example, the query presented in Figure 7 search for an instance of *skos:Concept* that contains the French word "carotte" in their preferred French label.

As shown in Figure 7, there exist three *skos:Concept* instances that have a French preferred label containing the word "carotte." Remember that the expert should select FCU concept narrower than "*Usages_plantes_cultivees*". When clicking on one of the results of the SPARQL query displayed on Figure 7, e.g., *fcu:Carrottes_potageres*, the expert accesses the RDF description of the FCU concept, as depicted in Figure 3.

To find a taxon in TAXREF-LD, a text search interface is accessible at http://taxref.i3s.unice.fr/fct/. For example, looking for the text expression "*Daucus carota*" provides the results presented in Figure 8.

The experts should select the entity of the first line "taxref:taxon/94503." The prefix "taxref" indicates that the entity belongs to TAXREF-LD. The URL part "taxon" indicates that the entity represents a taxon in TAXREF-LD, that is to say an *OWL class*. Then, the experts click on the Web link to display more information about the entity, as shown in Figures 9, 10. By navigating through the Web interface, the experts follow the property "*has reference name* " to find its scientific name, as presented in Figure 11.

2.2.3. Curation rules

Three curation rules were created to document our mappings. The goal of curation rules is to prove that the source of information contains the mapping and can be used as mapping justification. Along with the rule, we indicate how to find the relevant elements in each of the mapped resources.

2.2.3.1. CR_Geves

The *Official Catalog of Species and Varieties of Cultivated Crops in France* from GEVES is a good source of information to find which crop usage is associated with a taxon. As shown in Figure 12, this catalog indicates for each cultivar (seed used by farmers):

- Its vernacular name in the field *Common species*,
- Its crop usage in the field *Category*,
- Its scientific name in the field *Botanical species*.

Notably, most of the time, the taxon rank indicated in this catalog is a species. First, the most specific crop usage was selected in the FCU branch "*Usages_plantes_cultivees*" by using the information from fields *Common species* and *Category*. Second, the taxon with the scientific name indicated in the field *Botanical species* is searched in TAXREF-LD. If it is possible to find the crop usage and the taxon without ambiguity, the mapping is created and its confidence value is fixed to one.

For example, based on the information provided in Figure 12 about a carrot cultivar, the expert should find in FCU thesaurus the instance of *skos:Concept fcu:Carottes_potageres* (see Figure 7). The expert should also find in TAXREF-LD the instance of *skos:Concept* identified by *taxref:name/94503* (see Figure 11) and the *OWL class* identified by *taxref:taxon/94503* (see Figure 9). Thus, two mappings are created with a confidence value of one:

- One between the *fcu:Carottes_potageres* crop usage and the *taxref:taxon/94503* species taxon using the *ontofcu:hasReferenceTaxon* annotation property.
- One between the crop usage *fcu:Carottes_potageres* and the *taxref:name/94503* reference scientific name using the *ontofcu:hasReferenceScientificName* object property.

2.2.3.2. CR_C3PO_KB

The Crop Planning and Production Process Ontology and Knowledge Base (C3PO KB) is a knowledge graph created by the Elzeard Enterprise (Darnala et al., 2021, 2022). This KG is another reference source about vegetable.[11] The knowledge graph is accessible as several TTL files on a git repository.[12] To create a mapping between a crop usage and a taxon, the experts should search into the TTL file related to the plant module.[13] Figure 13 presents an excerpt of this file. The vegetable description contains a crop usage indicated by the property *c3poplant:hasFCUTaxon* and a TAXREF-LD scientific name indicated by the property *c3poplant:hasScientificName*. The experts should find the corresponding taxon by searching in TAXREF-LD Web interface (following the link "*is the reference name of*"). Based on those information, some new mappings are created with a confidence value of one.

For example, Figure 13 presents the RDF description in TTL format of the instance of *c3poplant:VegetablePlant* related to carrot. Based on this description, the expert can construct two mappings with a confidence value of one as follows:

- One between the *fcu:Carottes_potageres* crop usage and the *taxref:taxon/133744* subspecies taxon using the *ontofcu:hasReferenceTaxon* annotation property.
- One between the crop usage *fcu:Carottes_potageres* and the *taxref:name/133744* reference scientific name using the *ontofcu:hasReferenceScientificName* object property.

2.2.3.3. CR_Experts

The experts can also use their own knowledge to state that a scientific name is linked to a given crop usage. In this case, the taxon name is searched in TAXREF-LD. If the scientific name is retrieved

[10] http://ontology.inrae.fr/frenchcropusage/sparql

[11] The ontology is accessible on AgroPortal repository https://agroportal.lirmm.fr/ontologies/C3PO/?p=summary.

[12] https://gitlab.com/serre-des-savoirs/c3po-kb

[13] The file is untitled "cultivatedplants-infos-result-triples.ttl" and is accessible at https://gitlab.com/serre-des-savoirs/c3po-kb/-/blob/main/plants/rdf-generated%20triples/cultivatedplants-infos-result-triples.ttl.

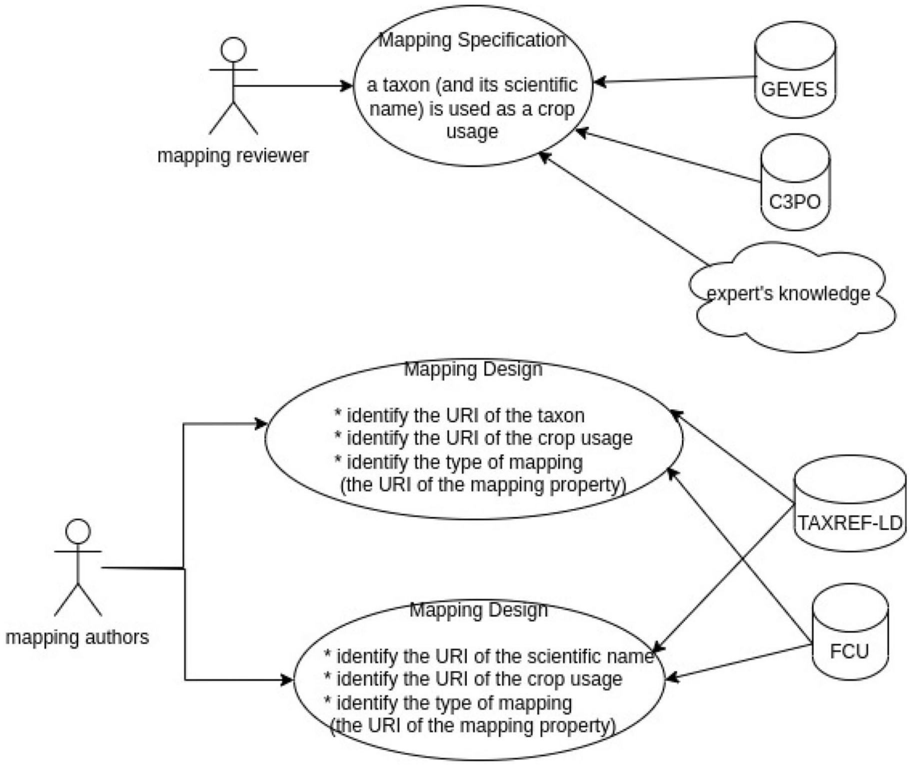
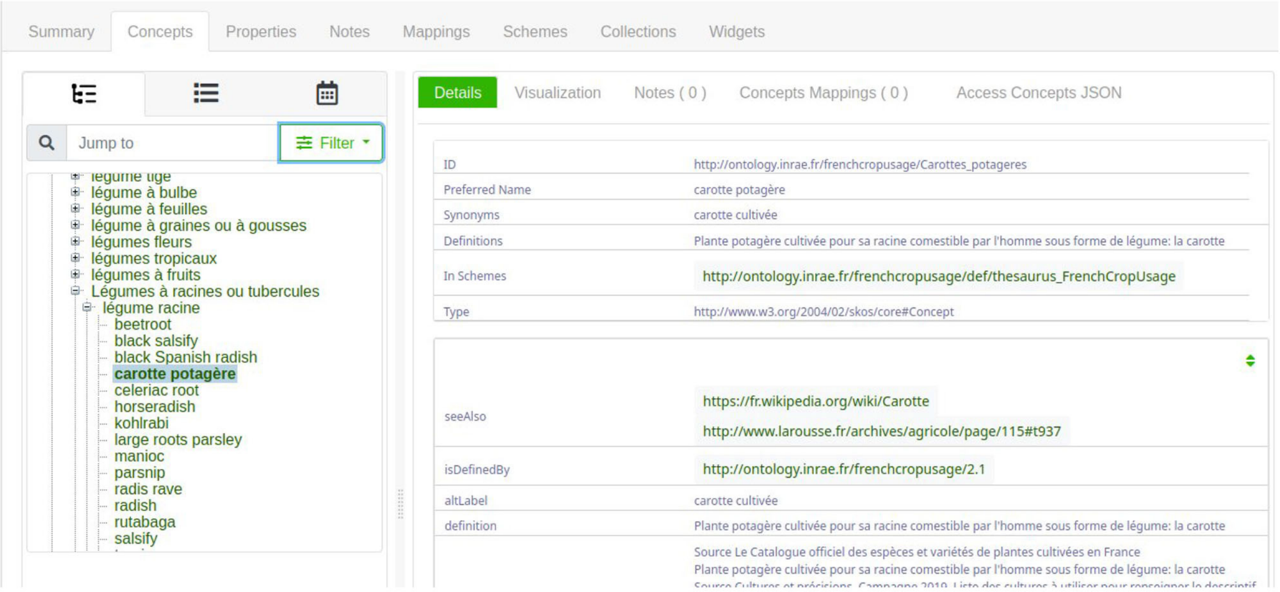

Mapping Specification
a taxon (and its scientific name) is used as a crop usage
mapping reviewer
GEVES
C3PO
expert's knowledge
Mapping Design
* identify the URI of the taxon
* identify the URI of the crop usage
* identify the type of mapping
(the URI of the mapping property)
TAXREF-LD
mapping authors
FCU
Mapping Design
* identify the URI of the scientific name
* identify the URI of the crop usage
* identify the type of mapping
(the URI of the mapping property)

FIGURE 5
The overview of the method.



légume tige
légume à bulbe
légume à feuilles
légume à graines ou à gousses
légumes fleurs
légumes tropicaux
légumes à fruits
Légumes à racines ou tubercules
- légume racine
  - beetroot
  - black salsify
  - black Spanish radish
  - carotte potagère
  - celeriac root
  - horseradish
  - kohlrabi
  - large roots parsley
  - manioc
  - parsnip
  - radis rave
  - radish
  - rutabaga
  - salsify

ID	http://ontology.inrae.fr/frenchcropusage/Carottes_potageres
Preferred Name	carotte potagère
Synonyms	carotte cultivée
Definitions	Plante potagère cultivée pour sa racine comestible par l'homme sous forme de légume: la carotte
In Schemes	http://ontology.inrae.fr/frenchcropusage/def/thesaurus_FrenchCropUsage
Type	http://www.w3.org/2004/02/skos/core#Concept

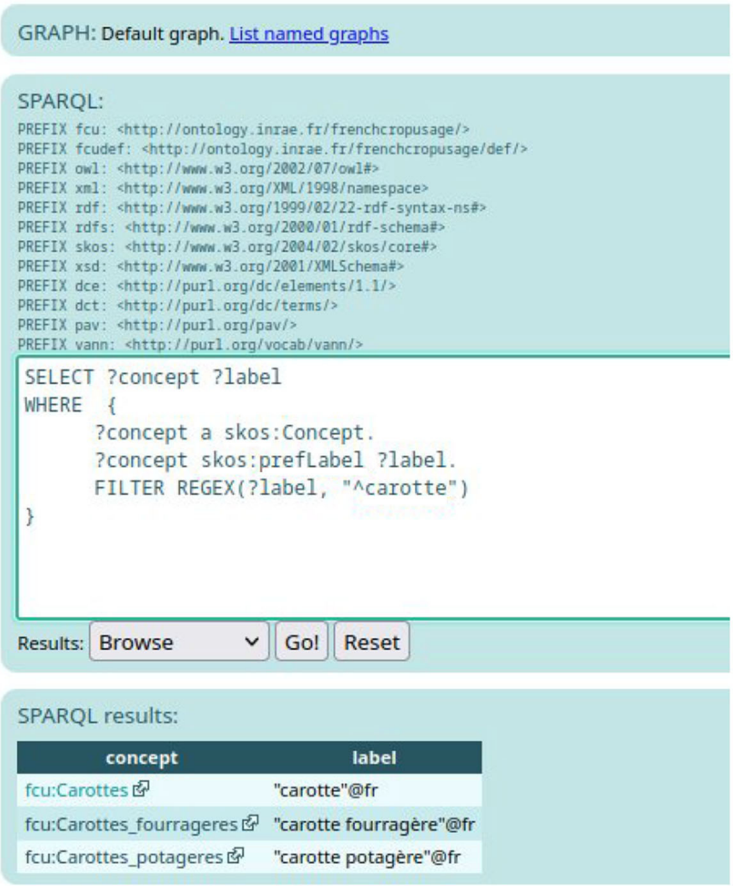
seeAlso	https://fr.wikipedia.org/wiki/Carotte http://www.larousse.fr/archives/agricole/page/115#t937
isDefinedBy	http://ontology.inrae.fr/frenchcropusage/2.1
altLabel	carotte cultivée
definition	Plante potagère cultivée pour sa racine comestible par l'homme sous forme de légume: la carotte
	Source Le Catalogue officiel des espèces et variétés de plantes cultivées en France Plante potagère cultivée pour sa racine comestible par l'homme sous forme de légume: la carotte Source Cultures et précisions. Campagne 2019. Liste des cultures à utiliser pour renseigner le descriptif

FIGURE 6
The AgroPortal concept tab interface showing the vegetable carrot crop usage.

and this name is the reference name for the taxon, two mappings are created as follows:

- One between the crop usage and the taxon using the *ontofcu:hasReferenceTaxon* annotation property.
- One between the crop usage and the scientific name using the *ontofcu:hasReferenceScientificName* object property.

If the scientific name is retrieved but the name is not the reference name for the taxon but a synonym name, three mappings are created as follows:



```
PREFIX fcu: <http://ontology.inrae.fr/frenchcropusage/>
PREFIX fcudef: <http://ontology.inrae.fr/frenchcropusage/def/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xml: <http://www.w3.org/XML/1998/namespace>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dce: <http://purl.org/dc/elements/1.1/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX pav: <http://purl.org/pav/>
PREFIX vann: <http://purl.org/vocab/vann/>
SELECT ?concept ?label
WHERE  {
      ?concept a skos:Concept.
      ?concept skos:prefLabel ?label.
      FILTER REGEX(?label, "^carotte")
}
```

concept	label
fcu:Carottes	"carotte"@fr
fcu:Carottes_fourrageres	"carotte fourragère"@fr
fcu:Carottes_potageres	"carotte potagère"@fr

FIGURE 7
The results of the SPARQL query related to carrot.

- One between the crop usage and the taxon using the *ontofcu:hasReferenceTaxon* annotation property.
- One between the crop usage and the reference scientific name using the *ontofcu:hasReferenceScientificName* object property.
- One between the crop usage and the synonym name using the *ontofcu:hasSynonymousScientificName* object property.

If the scientific name is not retrieved (for example, "*Cichorium intybus var. foliosum Hegi f. cylindricum*" is the scientific name of a form taxon), but the name of a parent taxon can be retrieved ("*Cichorium intybus var. foliosum Hegi* " is the scientific name of the variety taxon), the experts should restart the process by looking for the reference scientific name of the parent taxon. This implies that the properties used will be *ontofcu:hasGenericTaxon* and *ontofcu:hasGenericScientificName*.

The confidence value all these mappings is fixed to one.

3. Analysis

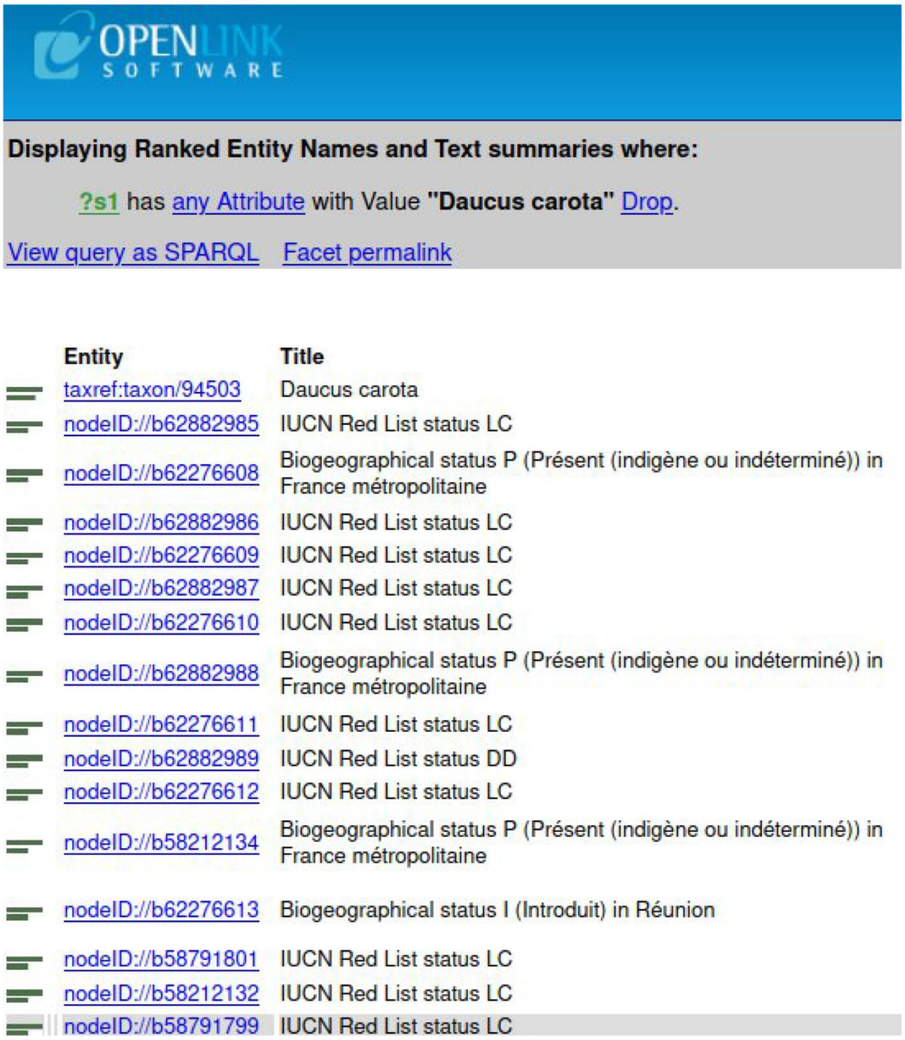
3.1. Alignment challenges

Although the taxonomies and lists of cultivated plants (or crop usages) refer to living organisms, their alignment raises several challenges that require manual curation in practice.

A taxonomy is a hierarchical structure that presents a set of hypotheses uttered by taxonomists on taxa and their relationships. The names of taxa are regulated by written conventions recorded in nomenclature codes. The current code in force for plants is the *International Code of Nomenclature for algae, fungi, and plants*, also known as the *Shenzhen Code* (Zhang et al., 2013). By their very nature, taxonomies change over time as the scientific consensus between taxonomists evolves. Taxa may merge, split, spawn, or change taxonomic rank (e.g., from *species* to *subspecies*). Nomenclature codes acknowledge this fact and were designed to stabilize taxon names as much as possible, given the versatility of taxonomies.

At a given time, a taxon has a single valid name and possibly a list of synonyms that have spawned from previous studies. Changes in the nomenclature are published on an *ad hoc* basis in the scientific literature. However, taxonomies and lists of cultivated plants are updated at a different pace. As a consequence, different references might not be completely up-to-date and may exhibit discrepancies on specific taxa. For instance, a list of cultivated plants may use a scientific name which is no longer valid or whose taxon has changed parent taxon or rank.

Additionally, nomenclature codes apply unambiguously to the species and subspecies taxon ranks but not to the lower ranks (e.g., *variety* or *cultivar*). On the other hand, the naming of cultivated



Displaying Ranked Entity Names and Text summaries where:

?s1 has any Attribute with Value **"Daucus carota"** Drop.

View query as SPARQL Facet permalink

Entity	Title
taxref:taxon/94503	Daucus carota
nodeID://b62882985	IUCN Red List status LC
nodeID://b62276608	Biogeographical status P (Présent (indigène ou indéterminé)) in France métropolitaine
nodeID://b62882986	IUCN Red List status LC
nodeID://b62276609	IUCN Red List status LC
nodeID://b62882987	IUCN Red List status LC
nodeID://b62276610	IUCN Red List status LC
nodeID://b62882988	Biogeographical status P (Présent (indigène ou indéterminé)) in France métropolitaine
nodeID://b62276611	IUCN Red List status LC
nodeID://b62882989	IUCN Red List status DD
nodeID://b62276612	IUCN Red List status LC
nodeID://b58212134	Biogeographical status P (Présent (indigène ou indéterminé)) in France métropolitaine
nodeID://b62276613	Biogeographical status I (Introduit) in Réunion
nodeID://b58791801	IUCN Red List status LC
nodeID://b58212132	IUCN Red List status LC
nodeID://b58791799	IUCN Red List status LC

FIGURE 8
The results of the text query "*Daucus carota*" in TAXREF-LD.

plants is not regulated in any way, and a cultivated plant name is not even required to refer to a scientific taxon.

Furthermore, due to their complexity, nomenclature codes are not always strictly observed. For instance, the *Official Catalog of Species and Varieties of Cultivated Crops in France* gives the authority without the date (e.g., "L." instead of "L. 1758"). Lists of cultivated plants are often built by aggregating several primary sources. However, these resources seldom cite primary sources, hindering the assessment of the confidence of their content.

Finally, vernacular names used to denote cultivated plants are sometimes specific to a given region, making alignment locale-dependent.

There are more technical difficulties related to modeling choices for representing taxa, names, and cultivated plants. For example, TAXREF-LD (Michel et al., 2017) strictly separates taxonomy and nomenclature. Other resources do distinguish taxonomy and nomenclature, representing both taxa and their names at the same level. Some classifications represent only scientific names, such as the Catalog of Life (Hobern et al., 2021). The lists of cultivated plants often retain only a scientific name instead of a taxon, a name which may no longer be valid.

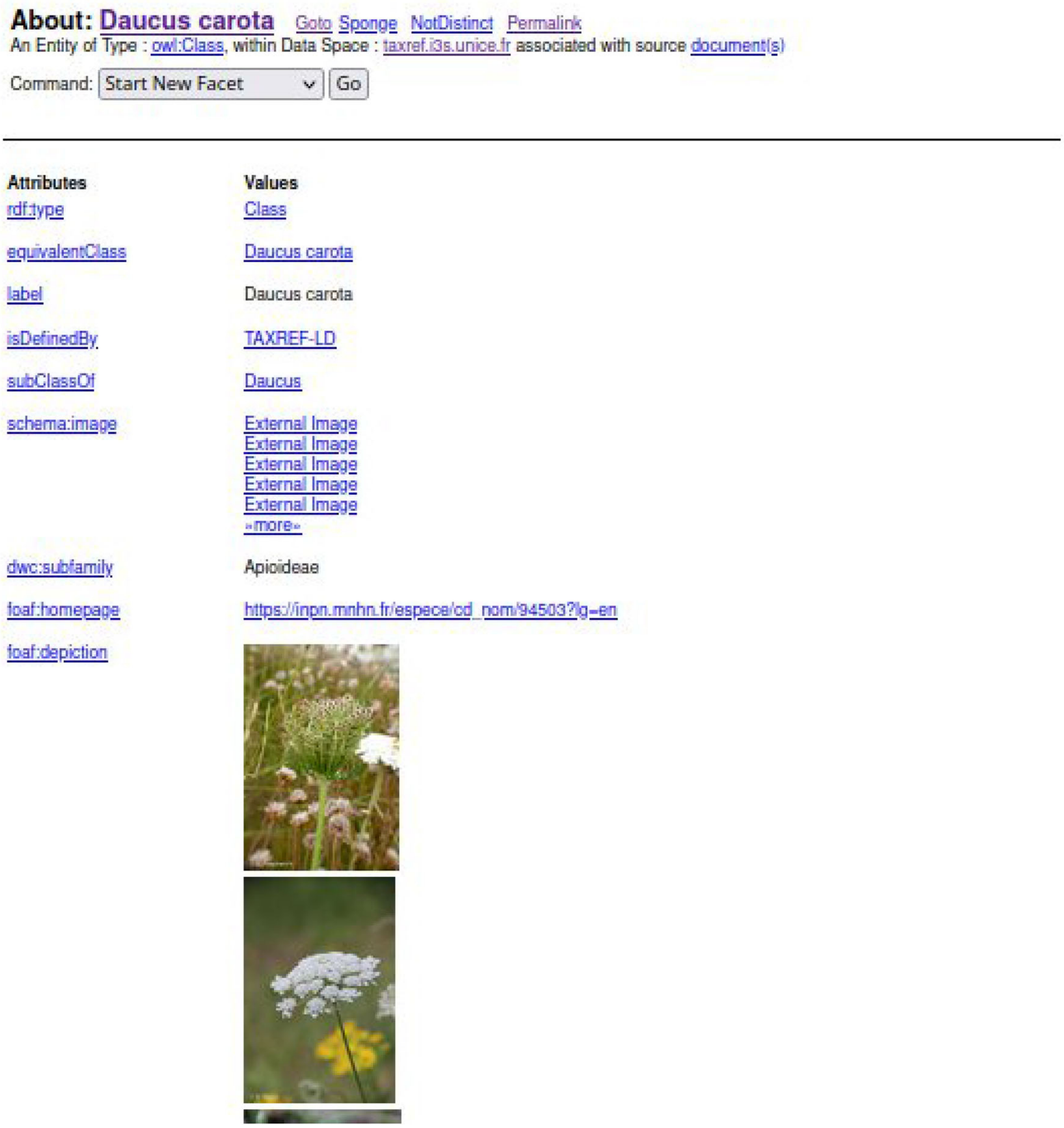
These design and modeling variations raise recurring questions on the type of objects to be aligned: do we align two taxa, or a taxon and a name, or a cultivated plant and a taxon, etc.?

Facing these challenges, we have proposed to define new mapping properties to represent the link between a crop usage, a taxon, and its scientific name.

3.2. SSSOM model analysis

SSSOM is a Simple Standard for Sharing Ontological Mappings which provides means to represent rich metadata for mappings and mapping sets. This standard representation has several advantages as follows:

- Entities of different nature can be mapped, for example, an instance of *owl:Class* with an instance of *skos:Concept*.



About: Daucus carota Goto Sponge NotDistinct Permalink

An Entity of Type : owl:Class, within Data Space : taxref.i3s.unice.fr associated with source document(s)

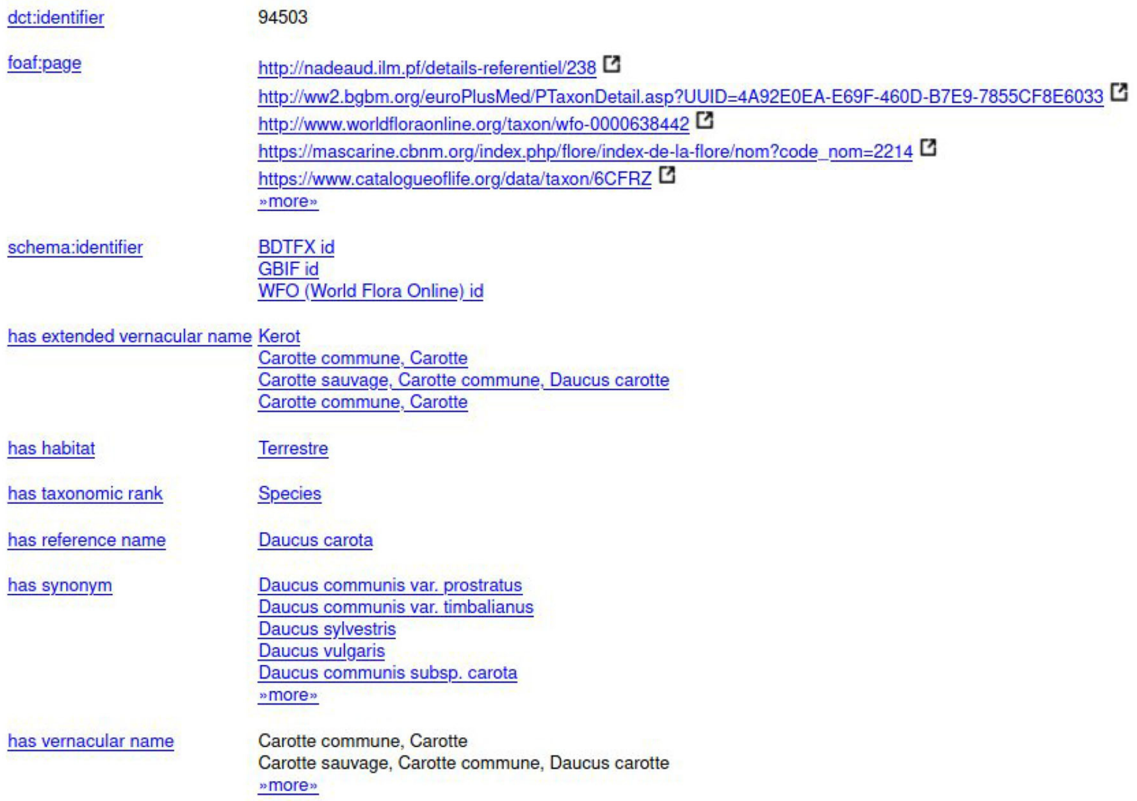
Command: Start New Facet Go

Attributes	Values
rdf:type	Class
equivalentClass	Daucus carota
label	Daucus carota
isDefinedBy	TAXREF-LD
subClassOf	Daucus
schema:image	External Image External Image External Image External Image External Image »more»
dwc:subfamily	Apioideae
foaf:homepage	https://inpn.mnhn.fr/espece/cd_nom/94503?lg=en
foaf:depiction	

FIGURE 9
First page of the description of taxon "*Daucus carota*" in TAXREF-LD.

- SSSOM does not impose the use of specific mapping properties such as *owl:equivalentClass*, the *oboInOwl hasDbXref*, or the well-known SKOS mapping properties, e.g., *skos:exactMatch*. In the context of this study, we were, thus, able to define the mapping properties relevant to our application (see Section 2.1.3).
- A set of mappings can become a dataset to be stored and shared independently of the aligned resources. Provenance metadata at dataset level allow to document the context in which the mappings are valid, for example, for the needs of an application. It is also possible to publish successive versions of a mapping set in a transparent way.
- Mappings can be described as first class objects and be referenced individually with an identifier (URI). This allows to declare equivalences between mappings coming from various sources to build aggregated mappings.
- Authors and reviewers of the mappings can be credited for their study as SSSOM recommends the use of ORCIDs with the author_id and reviewer_id properties.

The SSSOM project provides a rich documentation with many examples which facilitates the study. In addition, the SSSOM community is very active in working on improving and extending the model based on feedback from use cases issued from varied communities. This study of publishing mappings led us to



dct:identifier	94503
foaf:page	http://nadeaud.ilm.pf/details-referentiel/238 http://ww2.bgbm.org/euroPlusMed/PTaxonDetail.asp?UUID=4A92E0EA-E69F-460D-B7E9-7855CF8E6033 http://www.worldfloraonline.org/taxon/wfo-0000638442 https://mascarine.cbnm.org/index.php/flore/index-de-la-flore/nom?code_nom=2214 https://www.catalogueoflife.org/data/taxon/6CFRZ »more»
schema:identifier	BDTFX id GBIF id WFO (World Flora Online) id
has extended vernacular name	Kerot Carotte commune, Carotte Carotte sauvage, Carotte commune, Daucus carotte Carotte commune, Carotte
has habitat	Terrestre
has taxonomic rank	Species
has reference name	Daucus carota
has synonym	Daucus communis var. prostratus Daucus communis var. timbalianus Daucus sylvestris Daucus vulgaris Daucus communis subsp. carota »more»
has vernacular name	Carotte commune, Carotte Carotte sauvage, Carotte commune, Daucus carotte »more»

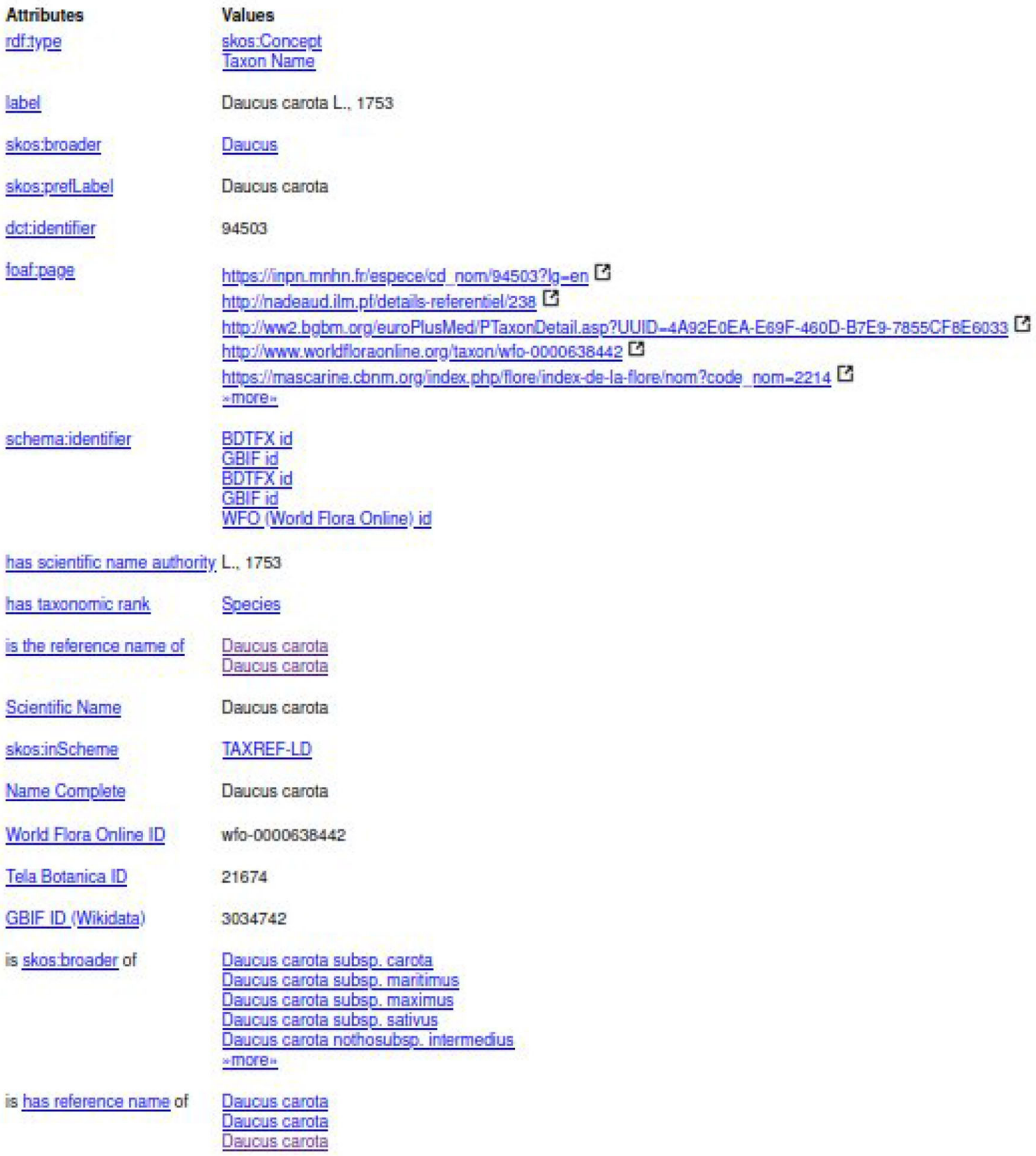
FIGURE 10
Second page of the description of taxon "*Daucus carota*" in TAXREF-LD.

contribute to the discussions taking place on the SSSOM github repository.[14] We expressed our need to document the rules that led us to assess a given mapping and discussed with the SSSOM authors on how to represent and expose these rules. As this new feature was plebiscited by several users, two new properties *sssom:curation_rule* and *sssom:curation_rule_text* were recently added to the SSSOM model. The rule can be represented either directly in the mapping set (using sssom:curation_rule_text) or be published as a RDF resource and be referenced with its URI in the mapping set (using sssom:curation_rule). As we work in TSV, we decided to give a short version of each curation rule in the TSV file of the mapping set, to refer to this article for a much finer description of our curation rules 2.2.3.

SSSOM, however, still shows some limitations or elements that are not fully mature as follows:

- Complex mappings: In our case, a crop usage may be fulfilled by a combination of taxa; for instance in viticulture, a vine plant can be composed of a rootstock and a graft. The rootstock is the buried part of the vine and serves as a support for the graft. Here, we may want to align *crop A* with a combination of a *taxon B* (rootstock) and a *taxon C* (graft) and indicate their respective roles. The future extension of SSSOM to handle complex mappings may include a property to assert the role of each component of a combination. If so, our RDF vocabulary would be extended to provide the types of roles in our specific context.
- Negative mappings: the open world assumption holds in Semantic Web models. This means that the absence of mapping does not mean that this does not exist or is incorrect. Moreover, we want to declare that some mappings are false or irrelevant. The SSSOM community investigated two solutions: adding a modifier column or creating negative mapping properties. As we developed our own mapping vocabulary, we decided to create some specific mapping properties to indicate that a mapping should not exist between FCU and TAXREF-LD: *ontofcu:hasInvalidTaxon / ontofcu:hasInvalidUsage / ontofcu:hasInvalidScientificName / hasInvalidVernacularName*.
- Confidence is defined in SSSOM as "A score between 0 and 1 to denote the confidence or probability that the match is correct, where 1 denotes total confidence." We are wandering if this property is relevant in the case of manual alignment. Concretely, the experts working on this mapping set faced two difficulties in filling out this field. First, it was difficult to fix a confidence value when they were not completely sure about a mapping, or even worse, when they disagreed. For this reason, and also because the SSSOM property lacks some clarity and is actually under discussion, we agreed

[14] https://github.com/mapping-commons/sssom/issues



Attributes	Values
rdf:type	skos:Concept Taxon Name
label	Daucus carota L., 1753
skos:broader	Daucus
skos:prefLabel	Daucus carota
dct:identifier	94503
foaf:page	https://inpn.mnhn.fr/espece/cd_nom/94503?lg=en http://nadeaud.ilm.pf/details-referentiel/238 http://ww2.bgbm.org/euroPlusMed/PTaxonDetail.asp?UUID=4A92E0EA-E69F-460D-B7E9-7855CF8E6033 http://www.worldfloraonline.org/taxon/wfo-0000638442 https://mascarine.cbnm.org/index.php/flore/index-de-la-flore/nom?code_nom=2214 »more»
schema:identifier	BDTFX id GBIF id BDTFX id GBIF id WFO (World Flora Online) id
has scientific name authority	L., 1753
has taxonomic rank	Species
is the reference name of	Daucus carota Daucus carota
Scientific Name	Daucus carota
skos:inScheme	TAXREF-LD
Name Complete	Daucus carota
World Flora Online ID	wfo-0000638442
Tela Botanica ID	21674
GBIF ID (Wikidata)	3034742
is skos:broader of	Daucus carota subsp. carota Daucus carota subsp. maritimus Daucus carota subsp. maximus Daucus carota subsp. sativus Daucus carota nothosubsp. intermedius »more»
is has reference name of	Daucus carota Daucus carota Daucus carota

FIGURE 11
Description of scientific name "*Daucus carota* L., 1753" in TAXREF-LD given as a whole by *rdfs:label* and split into its binomial name "*Daucus carota*" (*skos:prefLabel*) and the authority "L., 1753" (*txrfp:hasAuthority* whose label *has scientific name authority*).

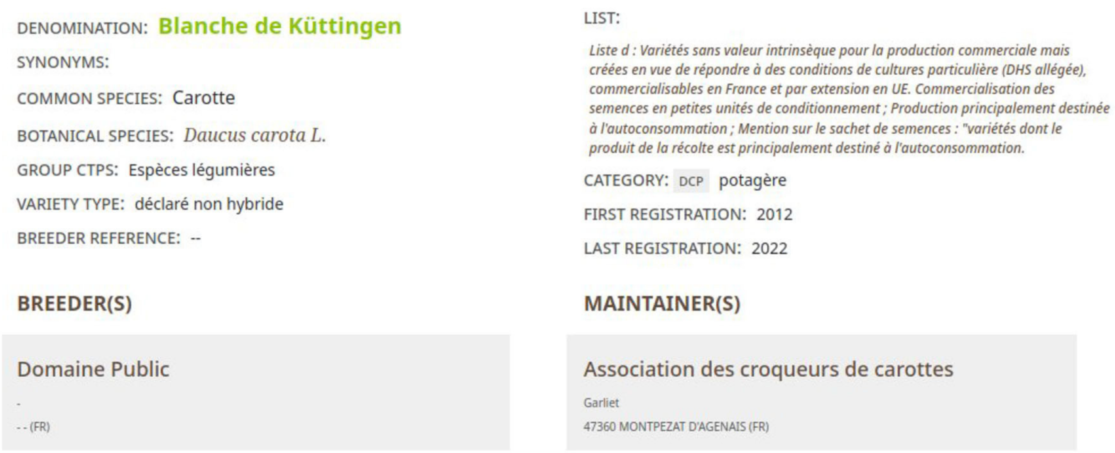
on publishing only mappings with a confidence value equal to one.

- Cardinality: as recommended in the guidelines, this property should be automatically filled in. It took a long time to compute the values and would be difficult to maintain when more mappings are added to the mapping set.

Overall, SSSOM is a rising metadata standard for sharing, analyzing, and integrating mappings. It covers our needs pretty well. The SSSOM project also offers a forum to discuss solutions with experts and practitioners from various domains.

4. Conclusion and perspectives

This article describes our work on the alignment of two complementary knowledge graphs useful in agriculture: the crop usage defined in the thesaurus of cultivated plants in France named French Crop Usage (FCU) and the taxa and associated scientific names defined in the French national taxonomic repository TAXREF for fauna, flora, and fungi. Due to the fact that automatic alignment methods provide poor results, a group of agricultural experts has produced a set of valid mappings between crop usages, taxa, and associated scientific names. To do so, a new RDF



DENOMINATION: Blanche de Küttingen

SYNONYMS:

COMMON SPECIES: Carotte

BOTANICAL SPECIES: *Daucus carota L.*

GROUP CTPS: Espèces légumières

VARIETY TYPE: déclaré non hybride

BREEDER REFERENCE: --

LIST:

Liste d : Variétés sans valeur intrinsèque pour la production commerciale mais créées en vue de répondre à des conditions de cultures particulière (DHS allégée), commercialisables en France et par extension en UE. Commercialisation des semences en petites unités de conditionnement ; Production principalement destinée à l'autoconsommation ; Mention sur le sachet de semences : "variétés dont le produit de la récolte est principalement destiné à l'autoconsommation.

CATEGORY: DCP potagère

FIRST REGISTRATION: 2012

LAST REGISTRATION: 2022

BREEDER(S)

Domaine Public

-

- - (FR)

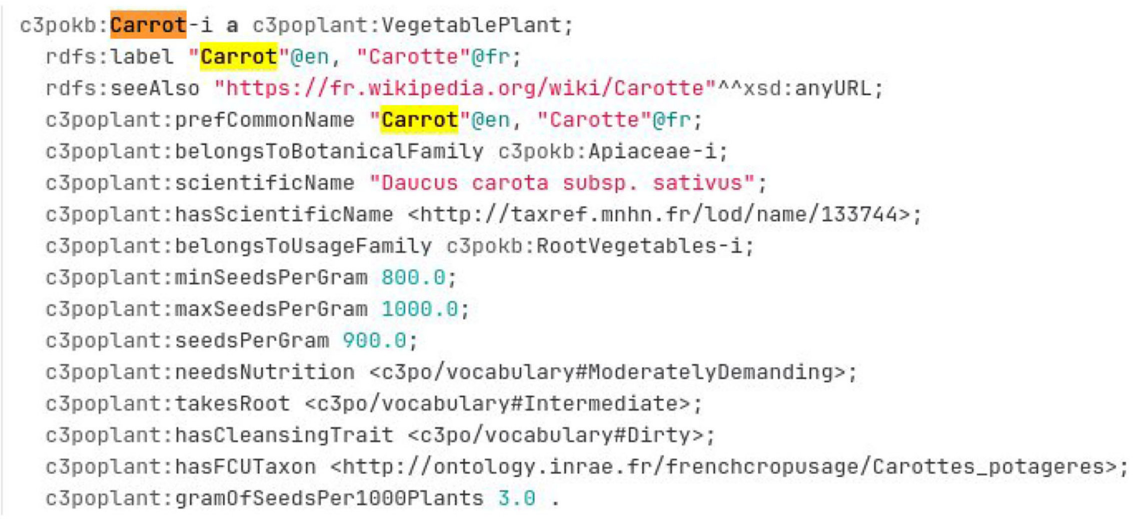
MAINTAINER(S)

Association des croqueurs de carottes

Garliet

47360 MONTPEZAT D'AGENAIS (FR)

FIGURE 12
The description of the carrot cultivar "*Blanche de Küttingen*" in the GEVES Catalog.



```
c3pokb:Carrot-i a c3poplant:VegetablePlant;
  rdfs:label "Carrot"@en, "Carotte"@fr;
  rdfs:seeAlso "https://fr.wikipedia.org/wiki/Carotte"^^xsd:anyURL;
  c3poplant:prefCommonName "Carrot"@en, "Carotte"@fr;
  c3poplant:belongsToBotanicalFamily c3pokb:Apiaceae-i;
  c3poplant:scientificName "Daucus carota subsp. sativus";
  c3poplant:hasScientificName <http://taxref.mnhn.fr/lod/name/133744>;
  c3poplant:belongsToUsageFamily c3pokb:RootVegetables-i;
  c3poplant:minSeedsPerGram 800.0;
  c3poplant:maxSeedsPerGram 1000.0;
  c3poplant:seedsPerGram 900.0;
  c3poplant:needsNutrition <c3po/vocabulary#ModeratelyDemanding>;
  c3poplant:takesRoot <c3po/vocabulary#Intermediate>;
  c3poplant:hasCleansingTrait <c3po/vocabulary#Dirty>;
  c3poplant:hasFCUTaxon <http://ontology.inrae.fr/frenchcropusage/Carottes_potageres>;
  c3poplant:gramOfSeedsPer1000Plants 3.0 .
```

FIGURE 13
The description of the carrot in the C3PO KB.

vocabulary of mapping properties was defined to align those plant descriptions. The metadata for the mappings and the mapping set are encoded with the Simple Standard for Sharing Ontological Mappings (SSSOM), a new model which offers means to report on the mapping provenance. To help the mapping creation, we provided some guidelines and tools to the experts. The produced mappings are available for download in Recherche Data Gouv, the federated national platform for research data in France.

Those mappings can be viewed as a first effort to test SSSOM Model using the TSV format. The mappings are manual and simple ones with high confidence value. Thus, they represent valid and consolidate mappings. We would like to enrich this mapping set by taking into account the whole Catalog of GEVES and C3PO KB. Both are evaluated as good source of information by our experts. We also plan to exploit these valid mappings to evaluate automatic alignment methods. The difficulty will be to manage the evolution of FCU and TAXREF-LD and keep up to date the mappings between those graphs. We would like to detect automatically obsolete mappings. This study is a first step to test the description of mappings produced by experts. We hope that it can help other use cases of mapping between complementary description defining different points of view on the same objects of study. We also hope that the provision of curated mappings will allow testing of new automatic methods capable of working with scientific names and vernacular names.

Data availability statement

The datasets presented in this study can be found in online repositories. TAXREF-LD: The version 15.2 of TAXREF-LD graph used for this study can be found in the AgroPortal repository https://agroportal.lirmm.fr/ontologies/TAXREF-LD. The github repository is https://github.com/frmichel/taxref-ld. The SPARQL EndPoint is https://taxref.mnhn.fr/sparql. FCU: The version 3.3 of the FCU thesaurus used for this study can be found in the AgroPortal repository: https://agroportal.lirmm.fr/ontologies/CROPUSAGE. The gitlab repository is https://gitlab.irstea.fr/copain/frenchcropusage. The SPARQL EndPoint is http://ontology.inrae.fr/frenchcropusage/sparql. SSSOM: The github repository is https://github.com/mapping-commons/sssom C3PO KB: The version 1.0 of the C3PO KB can be found in the gitlab repository https://gitlab.com/serre-des-savoirs/c3po-kb. The associated ontology can be found on the AgroPortal repository https://agroportal.lirmm.fr/ontologies/C3PO/?p=summary mapping set FCU TAXREF-LD. The mapping set between FCU and TAXREF-LD generated for this study can be found in the Research Data Gouv repository https://doi.org/10.57745/LVRFWJ the CSV file version presented in the article is available in article/Supplementary material.

Author contributions

SA and SB: SSSOM documentation. CF and FM: conceptualization and design of TAXREF-LD knowledge graph and associated research tool. CR and SB: conceptualization and design of FCU knowledge graph and associated research tool. FA and JR: conceptualization and design of C3PO knowledge graph. CR, FA, and JR: guidelines definition and test. CR and SA: writing—original draft preparation. CR, SA, CF, FM, and RB: writing—review and editing. CR: supervision. All authors contributed to the article and approved the submitted version.

Funding

This study was achieved with support from the Data to Knowledge in Agronomy and Biodiversity (D2KAB-www.d2kab.org) project that received funding from the French National Research Agency (ANR-18-CE23-0017). This study was also carried out by the French government as part of the Recovery Plan and the Future Investment Program “i-Nov”.

Conflict of interest

FA and JR were employed by Elzeard.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2023.1188036/full#supplementary-material

SUPPLEMENTARY DATA SHEET 1

SSSOM files that contains 63 mappings between vegetables or grapevine crop usage and their biological taxa.

References

Darnala, B., Amardeilh, F., Roussey, C., and Jonquet, C. (2021). “Crop Planning and Production Process Ontology (C3PO), a new model to assist diversified crop production,” in *IFOW 2021 - Integrated Food Ontology Workshop @ 12th International Conference on Biomedical Ontologies (ICBO)* (Bolzano, Italy).

Darnala, B., Amardeilh, F., Roussey, C., Todorov, K., and Jonquet, C. (2022). “Ontological representation of cultivated plants: linking botanical and agricultural usages,” in *MK 2022 - 1st Workshop on Modular Knowledge @ ESWC 2022 of CEUR Workshop Proceedings*, eds L. Bozzato, V. A. Carrier, T. Hahmann, and A. Zimmermann (Hersonissos, Greece), 165–173.

Gargominy, O., Tercerie, S., Régnier, C., Ramage, T., Dupont, P., Daszkiewicz, P., et al. (2021). *TAXREF v15, référentiel taxonomique pour la France: méthodologie, mise en uvre et diffusion*. Technical report.

Hobern, D., Barik, S.K., Christidis, L., Garnett, S. T., Kirk, P., Orrell, T. M. et al. (2021). Towards a global list of accepted species VI: The Catalogue of Life checklist. *Org. Divers. Evol.* 21, 677–690. doi: 10.1007/s13127-021-00516-w

Matentzoglu, 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. doi: 10.1093/database/baac035

Michel, F., Amardeilh, F., Bossy, R., Faron, C., Roussey, C., and Noûs, C. (2022). “Alignement entre sources : cas d’usage des plantes cultivées,” in *Journées francophones d’Ingénierie des Connaissances* (Saint-Étienne, France).

Michel, F., Gargominy, O., Tercerie, S., and Faron-Zucker, C. (2017). “A model to represent nomenclatural and taxonomic information as linked data. Application to the french taxonomic register, TAXREF,” in *Proceedings of the ISWC2017 workshop on Semantics for Biodiversity (S4BioDiv)* (Vienna, Austria: CEUR Workshop Proceedings). Available online at: http://ceur-ws.org/Vol-1933/paper-3.pdf (accessed September 01, 2023).

Miles, A., and Bechhofer, S. (2009). *SKOS Simple Knowledge Organization System Reference. W3C Recommendation*. World Wide Web Consortium, United States.

Zhang, N., Rossman, A., Seifert, K., Bennett, J., Cai, Y., Hillman, B., et al. (2013). *Impacts of the international code of nomenclature for algae, fungi, and plants (melbourne code) on the scientific names of plant pathogenic fungi*. APS Journal, Available online at: https://www.apsnet.org/edcenter/apsnetfeatures/Pages/Melbourne.aspx (accessed September 01, 2023).

frontiers | Frontiers in Artificial Intelligence

TYPE Methods
PUBLISHED 04 October 2023
DOI 10.3389/frai.2023.1209507

OPEN ACCESS

EDITED BY
Qichun Yang,
The University of Melbourne, Australia

REVIEWED BY
Shailesh Tripathi,
Tampere University of Technology, Finland
Zhen Liu,
The Hong Kong University of Science and Technology, China

*CORRESPONDENCE
Arjun Pakrashi
✉ arjun.pakrashi@ucd.ie

RECEIVED 20 April 2023
ACCEPTED 08 September 2023
PUBLISHED 04 October 2023

CITATION
Pakrashi A, Wallace D, Mac Namee B, Greene D and Guéret C (2023) CowMesh: a data-mesh architecture to unify dairy industry data for prediction and monitoring.
Front. Artif. Intell. 6:1209507.
doi: 10.3389/frai.2023.1209507

CowMesh: a data-mesh architecture to unify dairy industry data for prediction and monitoring

Arjun Pakrashi[1,2,3]*, Duncan Wallace[1,2,3], Brian Mac Namee[1,2,3], Derek Greene[1,2,3] and Christophe Guéret[4]

[1]School of Computer Science, University College Dublin, Dublin, Ireland, [2]Insight Centre for Data Analytics, Dublin, Ireland, [3]VistaMilk SFI Research Centre, Teagasc Moorepark, Fermoy, Ireland, [4]Accenture Labs, Dublin, Ireland

Dairy is an economically significant industry that caters to the huge demand for food products in people's lives. To remain profitable, farmers need to manage their farms and the health of the dairy cows in their herds. There are, however, many risks to cow health that can lead to significant challenges to dairy farm management and have the potential to lead to significant losses. Such risks include cow udder infections (i.e., mastitis) and cow lameness. As automation and data recording become more common in the agricultural sector, dairy farms are generating increasing amounts of data. Recently, these data are being used to generate insights into farm and cow health, where the objective is to help farmers manage the health and welfare of dairy cows and reduce losses from cow health issues. Despite the level of data generation on dairy farms, this information is often difficult to access due to a lack of a single, central organization to collect data from individual farms. The prospect of such an organization, however, raises questions about data ownership, with some farmers reluctant to share their farm data for privacy reasons. In this study, we describe a new *data mesh* architecture designed for the dairy industry that focuses on facilitating access to data from farms in a decentralized fashion. This has the benefit of keeping the ownership of data with dairy farmers while bringing data together by providing a common and uniform set of protocols. Furthermore, this architecture will allow secure access to the data by research groups and product development groups, who can plug in new projects and applications built across the data. No similar framework currently exists in the dairy industry, and such a data mesh can help industry stakeholders by bringing the dairy farms of a country together in a decentralized fashion. This not only helps farmers, dairy researchers, and product builders but also facilitates an overview of all dairy farms which can help governments to decide on regulations to improve the dairy industry at a national level.

KEYWORDS

data mesh, data fabric, digital agriculture, dairy industry, decentralized data

1. Introduction

The dairy industry is experiencing strong global growth (Douphrate et al., 2013; Bhat et al., 2022) accompanied by significant transformations through the increasing adoption of digital technologies (Borchers and Bewley, 2015; Gargiulo et al., 2018; Hansen et al., 2019; Gabriel and Gandorfer, 2023). As the industry enables this growth by producing dairy products more efficiently, it becomes essential to focus on cow health, as well as long-term factors, such as environmental impacts and profitability (Barkema et al., 2015; Bhat et al., 2022). The adoption of digital technologies on dairy farms (Gabriel and Gandorfer, 2023) means that considerably more data are being generated from dairy farms. These data are used to help farmers monitor their farms, make decisions, and achieve their goals around production, profitability, and cattle welfare and also to help administrative bodies to set national and international policies.

Additional availability of farm data also allows advanced statistical and machine learning techniques to be applied to support farm decision-making. Cow health has received significant attention in this regard. Early prediction of ailments in cows can reduce financial losses and improve cattle welfare. Recent adoption of technology in farm and data availability have triggered several advanced predictive analytic studies focused on dairy farms. For example, mastitis (an udder infection that afflicts cows) is one of the top reasons for monetary loss in dairy farms (Yalcin et al., 1999; Petrovski et al., 2006; Viguier et al., 2009), and several data-driven approaches to detecting mastitis are described in the literature (Ebrahimi et al., 2019; Bobbo et al., 2021; Ryan et al., 2021). There are also a number of studies that use data-driven approaches to detect lameness (Shahinfar et al., 2021; Altay and Albayrak Delialioğlu, 2022; Zhou et al., 2022) and ketosis (a metabolic disease in cows; Bauer and Jagusiak, 2022; Wang et al., 2023) using machine learning. As well as cow health, there are other aspects of dairy farming where data-centric advanced systems are being employed, including predicting herbage yield and composition (O'Hara et al., 2021; Albert et al., 2022), and estimating greenhouse gas emission (Chianese et al., 2009; Martin et al., 2017; Kadam and Vijayumar, 2018).

With the digitization of the dairy industry, research and product development are becoming more interdisciplinary. Sophisticated machine learning and statistical systems exploiting the data collected on farms are being developed by research groups in universities and organizations involving farmers, geneticists, computer scientists, and statisticians. A general trend is that several research groups perform research independently, with limited data sharing. This is often due to limited interoperability between data sources, data sharing challenges, and a lack of trust among stakeholders. Nonetheless, sharing the data generated on farms, as well as integrating different products (e.g., analysis, results, and services), has the potential to bring significant economic value to the agriculture industry (Wysel et al., 2021).

Wolfert et al. (2017) identified several major challenges in digital agriculture, which are also present in dairy farming: data ownership, data quality, sustainable integration of data sources, intelligence processing and analytics, business models, and openness of platforms. The World Economic Forum (WEF, 2023) is a non-profit organization that brings together global leaders to address critical issues and promote public-private cooperation, serving as a platform for networking, dialog, and shaping agendas for positive changes. The WEF summarizes the challenges in digital agriculture in three categories: fragmentation, standards, and access. A recent study by Fadul-Pacheco et al. (2022) describes data-oriented issues and demands of dairy farms. It was found that dairy farmers and non-dairy farmers (related to the dairy industry) believed that data sharing is important. However, issues with data ownership and data quality represented a significant area of concern. A significant portion was unsure about the chain of custody of data. Non-farmers (e.g., researchers and organizations without dairy farming expertise) were concerned about the lack of integration of data and, in some cases, not aware of the usefulness of data integration. The issue of trust when sharing data was raised in the study by Jakku et al. (2019), indicating that transparency, trust, and data ownership are major issues in sharing data. This highlights the need for a framework that ensures good data quality, effective data integration, transparent data use and ownership, and effective use of the data to build analytics and predictive systems. This, in turn, demonstrates that there is a requirement for a reliable data sharing framework in the dairy industry context which addresses the above issues.

In this study, we introduce a data mesh architecture, CowMesh, that addresses the challenges in data-driven dairy farming as described above. Specifically, we propose an architecture with a central semantic data product that provides interoperability among other data products and data domains by providing a high-level ontology of dairy farm components and a uniform data access protocol.

The rest of the study is structured as follows. Section 2 discusses related work. The proposed CowMesh architecture is described in detail in Section 3. A series of use cases from the Irish dairy farming context are then presented in Section 4, to show the typical usage of the CowMesh architecture. Key advantages and opportunities for the architecture are discussed in Section 5, and finally, Section 6 concludes the study.

2. Related work

According to WEF (2023), the key challenges around data in the agricultural domain can be summarized as follows:

- **Fragmentation**: Data are gathered from a variety of sources (sensors, satellites, etc.) and made available as different topical silos (soil data, seed data, etc.);
- **Standards**: There is no global standard or standardization body facilitating the expression of agricultural data;
- **Access**: Data need to be exchanged and connected in order to deliver value.

However, these challenges are actually not restricted to agriculture and are a larger concern across many different industry sectors. The proposal for making scientific data **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (FAIR; Wilkinson et al., 2016) is a pragmatic approach toward producing data in a better way. It has been designed with the scientific community in mind, but the principles are more globally applicable. FAIR pushes forward a number of core principles which align largely with the W3C Data on the Web Best Practices (DWBP)[1] but do not push forward any particular technology stack.

According to both FAIR and DWBP, vocabularies are a pillar of data publication. In the agricultural domain, the thesaurus

[1] https://www.w3.org/TR/dwbp/ (visited August 17, 2023).

AGROVOC[2] from the Food and Agriculture Organization is an important resource that is made available as a SKOS-based ontology (Caracciolo et al., 2013). This vocabulary is essential for describing agricultural concepts uniformly and in multiple languages. However, AGROVOC does not solve any of the access or fragmentation issues itself.

The Knowledge Graph "Agronomy Linked Data (AgroLD)" from the study by Larmande and Todorov (2021) (D2KB) is a good example of an integrated dataset. The content is FAIR-compliant and incorporates data coming from 15 different silos into a single, integrated dataset. The resulting dataset can be used to answer complex questions around plants and biology. The portal AgroLD[3] is the main entry point to explore the Knowledge Graph. In terms of an agriculture-related portal, and with a slightly different focus, LandPortal[4] serves integrated data about land use worldwide. The objective is to support queries around land ownership and arable land utilization.

We remark that, although vocabularies such as AGROVOC can support the creation of data portals aimed at particular needs, there is still a requirement for a more holistic approach. As outlined in the study by WEF (2023), there are a number of services needed around these portals in order to unlock their capabilities. It could also be interesting to consider creating a more flexible alternative to data portals constructed on a data aggregation approach. To achieve this aim, we propose the adoption of the recent and growing approach of the data to the agricultural domain (Joshi et al., 2021; Butte and Butte, 2022; Hooshmand et al., 2022; Bode et al., 2023; Dolhopolov et al., 2023; Goedegebuure et al., 2023; Pongpech, 2023).

3. The CowMesh architecture

In this section, we will introduce the CowMesh architecture by first providing a general overview of the Data Mesh approach and then explaining how this is adopted in our context. Finally, we describe the central Semantic Layer component, which is a Data Fabric, and its key role in CowMesh.

3.1. The data mesh

The data mesh concept was introduced by Dehghani (2022a,b) to define a set of principles for publishing data (Christ et al., 2022). From a technology point of view, a data mesh can be implemented using a variety of solutions and standards. The only particularity is the focus of application programming interface (API) to replace the (manual) handling of data dumps across systems. The novelty of the data mesh approach is not the data integration itself but rather how it is approached and considered from an organizational point of view. In particular, a data mesh is centered around four core concepts as follows:

- The preservation of **domain ownership** for the different domains may serve. As teams working together in a company system, the different data domains are expected to directly collaborate with each other and own their work—for example, in terms of data exchange or analytical work usage.
- The notion of **data products** replaces the older idea of data assets. The application of a product logic to data assets turns them into things that need to match a demand, whose value is assessed and production cost studied. Data domains may share datasets as product and/or analytical applications leveraging one or more other data products as a product of their own.
- A **data infrastructure platform** is put in place to let each data domain easily make data products available to the rest of an organization. This platform must not be limited to a particular domain and should facilitate both the creation and consumption of data products.
- An overall **federated governance** approach is applied to establish data standards and best practices to use the data mesh. This ensures the technical compatibility of all data products and can ensure compliance to rules and regulations.

It is interesting to note that these concepts describe one very well-known data publication platform, the World Wide Web. The Web features a strong notion of domain ownership. Each website publisher is responsible for its own websites and research communities publishing the outcomes of analytics on the Web, or the data within it, own those publications. Websites are by default treated as products and are routinely checked for view performances as well as optimized toward increasing those views. The web infrastructure platform based on a set of accessible software and programmatic tools make it possible for anyone to publish a new product on the Web. The W3C defines the standards and best practices that make the web run smoothly (HTTP, CSS, etc.). Finally, the Web often uses different attribution mechanisms such as Creative Common licensing[5] and document object identifiers[6] for attribution of content. Any new technical platform or data domain willing to join the mesh can easily do so as long as the compatibility with those standards is ensured.

3.2. Architecture

To tackle the previously discussed challenges, we propose an approach based on two emerging design patterns: a data fabric and a data mesh. Whereas, these two approaches can be described as opposing each other, especially in terms of data centralization and human versus process focus, we propose to combine the two patterns so that they complement each other. Our architecture, presented in Figure 1, is composed of:

- **Data Domains** such as research institutions, public institutions, and private actors. In the Irish context, these might be the Irish Cattle Breeding Federation (ICBF),[7]

[2] https://agrovoc.fao.org/browse/agrovoc/en/ (visited August 17, 2023).

[3] http://agrold.org (visited August 17, 2023).

[4] https://landportal.org/ (visited August 17, 2023).

[5] https://creativecommons.org/licenses/ (visited August 17, 2023).

[6] https://www.doi.org/ (visited August 17, 2023).

[7] https://www.icbf.com/ (visited August 17, 2023).

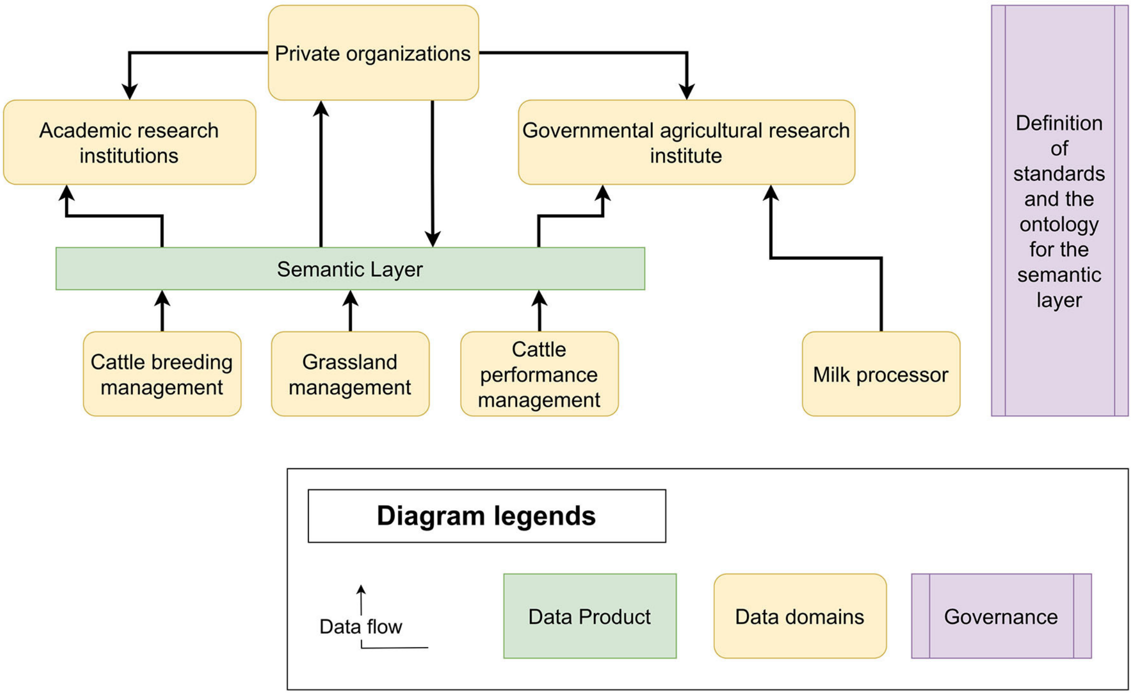

Private organizations
Academic research institutions
Governmental agricultural research institute
Definition of standards and the ontology for the semantic layer
Semantic Layer
Cattle breeding management
Grassland management
Cattle performance management
Milk processor
Diagram legends
Data flow
Data Product
Data domains
Governance

FIGURE 1
The CowMesh architecture shows the semantic layer (as a data product), several data domains, and a governance layer. A particularity of this architecture is the data product "Semantic Layer" which is a data fabric acting as a data infrastructure and interoperability layer for other data products.

Ornua,[8] or Teagasc.[9] Each of these stakeholders will have some datasets and tools that they contribute as data products (not depicted in Figure 1).

- **Data Products** are contributed by data domains. Examples in the Irish context include PastureBase[10] and the ICBF databases.[11]
- A **Semantic Layer** implemented as a data fabric is a core element of the proposed architecture and the only data product presented in Figure 1. The role of the semantic layer is to provide an integrated view over key data coming from different data domains. This does not prevent consumers of this data product from going back to the data domain sources in it—as, for example, illustrated in the link between milk processors and governmental research institutions (meaning, for instance, specific data access negotiations)—but does make it easier to consume the data.
- The **Governance** layer decides on the standards being used for the mesh overall, and the ontology driving the semantic layer.

It can be observed that the semantic layer is a data product lacking a defined data domain owning it. This is because we consider this as an open question left to specific implementations. In some cases, a research consortium might assume this role, while, in other cases, this would be one of the industrial stakeholders. For the specific Irish context, the research program VistaMilk[12] would assume this role.

In our architecture, the semantic layer is a data fabric, created as a data product and a central part of CowMesh. Since a key objective of a data mesh is to promote decentralization, having a central semantic layer might appear to be contradictory. However, its inclusion enhances the effectiveness of the data mesh by keeping the data decentralized while connecting them at a uniform semantic and conceptual level, thus establishing interoperability. Additionally, the semantic layer sets standards and provides governance to address data privacy, ownership, and access issues.

3.3. Semantic layer: a data fabric

The semantic layer data product is a data fabric in our architecture, which enables interoperability among different data domains and enables governance. Although this sits at the core of the architecture, and a data fabric is a centralized approach to manage data, this does not hinder the decentralized properties of the data mesh. Both data mesh and data fabric are architectural patterns to manage data in a distributed and complex environment. The main contrasting properties of the data mesh and data fabric are as follows:

[8] https://www.ornua.com/ (visited August 17, 2023).

[9] https://www.teagasc.ie/ (visited August 17, 2023).

[10] https://pasturebase.teagasc.ie (visited August 17, 2023).

[11] https://www.icbf.com/the-icbf-database/ (visited August 17, 2023).

[12] https://www.vistamilk.ie/ (visited August 17, 2023).

- **Scope**: A data fabric is typically designed to manage data across an entire organization, while a data mesh is more focused on specific domains, groups, or business units within an organization.
- **Approach**: A data fabric is more centralized in its approach, with a single unified architecture that connects and integrates data from different sources. On the other hand, a data mesh is more decentralized in its approach, with individual domains or teams responsible for managing their own data.
- **Governance**: A data fabric provides a centralized governance framework for managing data, while a data mesh relies on a decentralized governance model, where each domain or team is responsible for defining and enforcing its own governance policies.
- **Culture**: A data mesh is more focused on promoting a culture of data ownership and collaboration among teams, while a data fabric is more focused on standardization and consistency in the data management process.

Therefore, a data fabric is a centralized approach to managing data across an entire organization, while a data mesh is a more decentralized approach that focuses on promoting data ownership and collaboration within specific domains or business units. We use these contrasting properties to compliment each other to address the previously mentioned issues in dairy farms. The data fabric, used as a data product, establishes a uniform data model, access protocol, and governance model. On the other hand, the data mesh enables decentralized development of data products.

The role of the semantic layer in the CowMesh architecture involves:

- Providing an ontology of the concepts in the farm data that provides a uniform data model.
- Defining a set of protocols (through APIs) for accessing farm data through the concepts in the ontology.
- Integrating heterogeneous data domains in the CowMesh architecture to make them interoperable.

In summary, the semantic layer enables decentralized development in the data mesh while also providing interoperability, integration, and governance.

3.3.1. Ontology

Like any other domain, the data in dairy farming have a set of concepts related to them. For example, all data related to cows in a single herd or a collection of dairy farms can be observed as a concept "cow." The concept "cow," with respect to the data, is an abstract view of a cow which is described through its properties. A cow can be described by its unique identifier, date of birth, body weight, and various other attributes. Several concepts like this—such as farm, milk, and paddock—can also be defined. Different concepts will be interconnected to describe the abstract data representation of a farm. We propose an ontology for a dairy farm, which is shown in Figure 2.

An ontology such as the one shown in Figure 2 shows an example of how dairy industry data can be structured. The different concepts in our ontology are as follows:

- Cow: Describes the cow, date of birth, etc.
- Herd: Describes properties of a herd related with Cow.
- Soil: Properties of the farm's soil.
- Grass: Properties of the farm's grass.
- Paddock: Information about a farm's paddocks related to soil and grass.
- Milk Record: milk composition including fat, protein, lactose, urea, and somatic cell count.
- Daily Milk Record: Properties for daily milking, milk yield, milk flow etc.
- Milk: Describes overall milk.
- Farm: Information about the farm, e.g., location, area, nos. of herd, and nos. of cows per herd.
- Disease Occurrence: Record of occurrences of disease in cows such as clinical mastitis, sub-clinical mastitis, lameness, and respiratory disease.
- Treatment: Treatment performed for any disease.
- Farmer: Information about the farmer.

Such an ontology sits centrally in the semantic layer, abstracting the dairy farm data to the users. Some components of the ontology might describe sensitive or personal information, such as exact farm location or farmer information. The exact data to be shared and how it is shared will be controlled by the source and governed by the semantic layer according to the agreed privacy policy. Therefore, such data instances should be ideally anonymized, or omitted based on the policy, so as to preserve data privacy (e.g., the anonymization of data in the MilkMap system described later in Section 4.2). It is important to note that the semantic layer only enables interoperability by providing an abstract model of the data (ontology) and a protocol (though an API) for data access.

3.3.2. Access and access protocol

One of the main services of the semantic layer as a data product is to provide an access protocol for the farm data with respect to the provided ontology. All the data products in the CowMesh should be able to use this protocol to access data. This process can be implemented through an API, which refers to a set of protocols and tools that allows different software applications to communicate and interact with each other, thus enabling data and functionality sharing across systems. An API defines the methods and data structures that developers can use to integrate services or features into their own applications. This allows consumers to retrieve farm data in the CowMesh uniformly, even though the different data domains may have differently structured data and access protocols. The API is implemented in the semantic layer and is a data product, although the API can also reside in the source data domain, if required. For example, if a data domain is public, the data can be mapped by the semantic layer, and the API implementation can reside in the semantic layer. On the other hand, if the data domain is privately owned, the API implementation can reside as a data product with the private owner, but the access protocol and the compliance with the ontology are ensured. The particular implementation depends on the specific case. The semantic layer

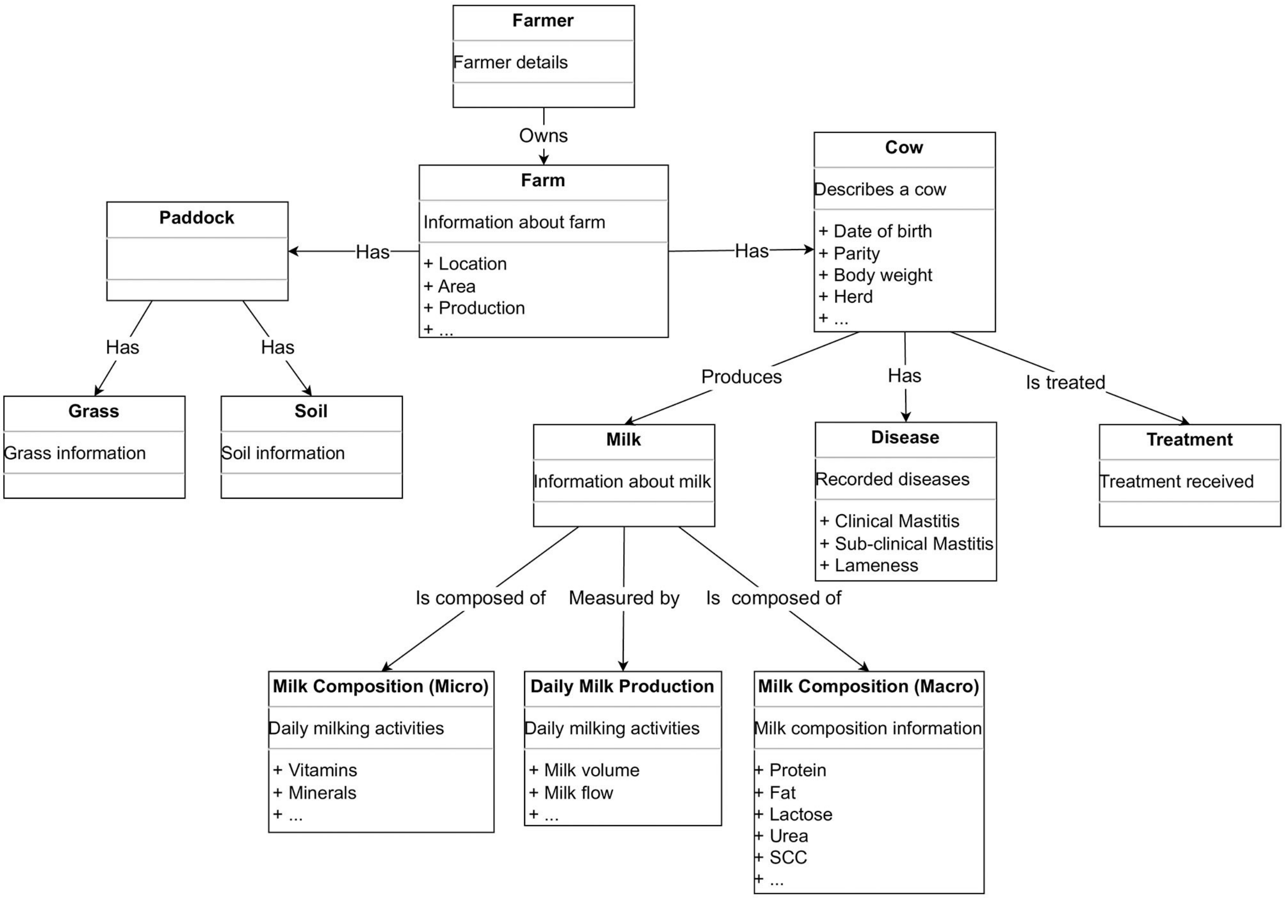


FIGURE 2
An ontology showing an example of how the concepts in the semantic layer of the CowMesh may be structured.

only provides the guidance through which it should be done, and specific implementation can be selected on a case-by-case basis.

One of the key features, here, can provide different degrees of access control and log data access, which helps keep track of the chain of custody of data. This not only makes the data access transparent but also provides an opportunity for monetization.

Another service which can be provided is to keep some commonly-used clean data in the semantic layer for easy access. This helps the products get the benefits of using pre-processed data from a central source. That said, this does not bind the data product to the limitations of the central data source, as the data products can always use the same data from the original source.

3.3.3. Data product integration

Enabling access to farm data through one uniform set of protocols requires that all data domains agree on the specified ontology. One of the key functions of the semantic layer data product is to add and update new data products in the CowMesh. To enable such integration of the data products, each product needs to provide access to their data in such a way that it complies with the data ontology and the access protocol set by the semantic layer. For example, cow milk property information may be stored in two data domains, one public and one private, which are stored in different formats. To integrate these two organizations into data domains in the CowMesh, the data products should be able to access them through protocols set by the semantic layer while providing the necessary access control and privacy.

The integration of data sources can be achieved by introducing the role of a Knowledge Scientist (KS; Fletcher et al., 2020), who will perform this integration. A KS is a person who represents a bridge between the underlying data and the business requirements. In our case, a KS would communicate with the two data domains, to understand the structure of the data which they are willing to make a part of the CowMesh. The KS will have a detailed understanding of the CowMesh and the semantic layer and the required knowledge of the data source through communication. This translation can be implemented at the semantic layer end or the data product end, which can be decided on a case-by-case basis and is implementation-specific.

It is important to emphasize that the KS does not need to be an expert in the dairy industry or know specific properties about the data or data cleaning. The key role of the KS is to have basic knowledge about the concepts related to dairy industry data, knowledge about CowMesh, the ontology, and the role of the semantic layer, ability to understand requirements from communications with the data domains, and understanding the related technologies of the specific implementation.

The data sources in the CowMesh do not necessarily need to represent a large public or private organization. Smaller and independent data hosts can also participate easily in the process by following the solid protocol,[13] guided by the protocols set by the semantic layer. A solid pod is a personal online data store that empowers users with control over their data, following the principles of the solid project for decentralized and secure data management on the web. The basic idea behind solid is to allow users to store their data in a "pod," which is a personal online data store that they control. Users can, then, grant access to their data to apps and services as they see fit, rather than having their data silo-ed in different apps and services controlled by large corporations. Solid aims to provide a more open, decentralized, and user-controlled web, where individuals have more autonomy over their personal data and are empowered to choose which apps and services they want to use and share their data with. Therefore, instead of having full infrastructure like an organization, an individual (e.g., a farmer) with data can contribute to the CowMesh by making their data accessible as data products through solid pods, which are easy to deploy. Services such as Inrupt[14] can be used to deploy a solid pod easily, with the assistance of a KS following the protocols set by the semantic layer.

3.3.4. Governance

The governance of data, data access, and ownership is simplified by the CowMesh architecture through the semantic layer. The data domains own their data, and they decide whether to keep the data within the semantic layer or rather to keep them on their private server. Each data domain and data product can have their own governance. However, to be a part of the CowMesh, the data domains, and data products need to conform to the standards and protocols set by the semantic layer. This provides two levels of governance. The data products in the data mesh as an independent unit may have their own governance. In addition, by being a part of the CowMesh, they fall under the uniform set of protocols and standards. This streamlines the governance of the entire CowMesh. Better governance will also encourage the opening of controlled channels from private organizations through CowMesh, which can help to facilitate greater collaboration between the dairy industry and academic researchers.

One open question revolves around who will govern the CowMesh. While the answer will be specific to the context in which CowMesh is being implemented, some possibilities are as follows: (1) one of the member organizations of the CowMesh can take responsibility for governance; (2) several members of the organizations of the CowMesh can form a governance forum; (3) a neutral organization can act as a governing body. One successful example of such a governing body is The Open Subsurface Data Universe (OSDU),[15] which regulates how oil and gas companies manage and analyze subsurface data. The goal of OSDU is to create a common data platform that allows the member companies to share and collaborate on subsurface data while also providing secure and scalable access to the data. OSDU is under the guidance of The Open Group,[16] a global consortium that brings together industry, government, and academia to develop open standards and best practices for technology.

4. Use case implementations

In this section, we will present two use cases which can benefit from the proposed architecture and overcome current challenges in data-driven dairy farming. Figure 3 shows an example of CowMesh architecture in the context of the Irish agricultural sector. Here, we extend Figure 1 to demonstrate data domains and data products, together with their interactions in Ireland. In this example, the semantic layer is owned by the VistaMilk data domain, and a data product "CowReport," which provides a periodic summary insight into the data from different sources. Such a report can directly help the farming industry analyze the data from a higher level perspective and help the stakeholders to take informed decisions.

Teagasc is the Agriculture and Food Development Authority in Ireland and has a data product PastureBase (Hanrahan et al., 2017) related to countrywide grassland management. The Irish Cattle Breeding Federation (ICBF) is a non-government organization that provides a large data repository for several areas related to dairy farms. Each academic institution or research center involved in agriculture research can be a data domain. For instance, we show some data products within the Insight Centre for Data Analytics[17] data domain. The possible products within this domain and how they interact are described in the following subsections.

4.1. Mastitis prediction

Mastitis is an inflammatory response of the udder in the cow's mammary gland caused due to microorganism infections. Mastitis is divided into two types, namely, (a) clinical mastitis, where symptoms are visible to the naked eye; and (b) sub-clinical mastitis, where the symptoms are not visible but can be measured though testing. Both of the variants compromise the health and wellbeing of the cows which results in negative impact on milk production volume and quality (Halasa et al., 2007), increased veterinary costs (Cavero et al., 2007), and an increased risk of culling. Mastitis is one of the most common infections on dairy farms globally, with ∼20–30% of cows in any herd likely to become infected annually (Heringstad et al., 2000). Therefore, the ability to predict the onset of clinical or sub-clinical mastitis in cows ahead of time will be of great benefit on dairy farms.

[13] https://solidproject.org/ (visited August 17, 2023).

[14] https://www.inrupt.com/ (visited August 17, 2023).

[15] https://osduforum.org/ (visited August 17, 2023).

[16] https://www.opengroup.org/ (visited August 17, 2023).

[17] https://www.insight-centre.org/ (visited September 19, 2023).

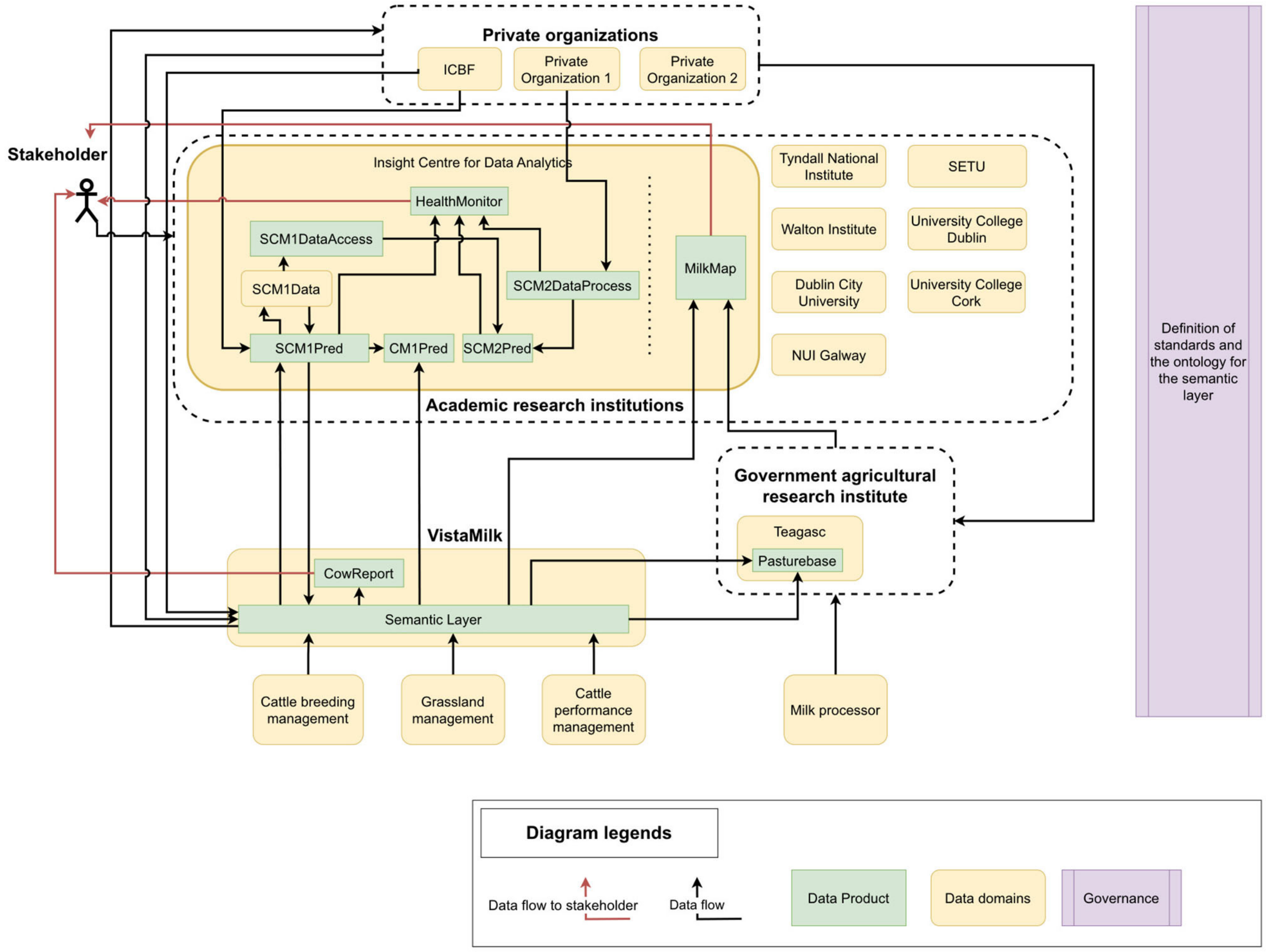


FIGURE 3
A CowMesh architecture implementation showing several data products for a specific use case.

4.1.1. Predicting mastitis: traditional process

Prediction of mastitis (both clinical and sub-clinical) has been previously performed using data-driven statistical and machine learning methods (Ebrahimi et al., 2019; Anglart et al., 2020; Bobbo et al., 2021). A recent study by Pakrashi et al. (2023) addressed this issue by predicting sub-clinical mastitis in Irish dairy cows up to 7 days ahead of time. Pakrashi et al. (2023) use machine learning algorithms to train prediction models using the data from seven Irish research farms spanning 9 years of data, which consisted of the following information:

- Daily milk yield and other milking information;
- Milk composition (e.g., fat, lactose, protein, and somatic cell count)
- General cow features available at the farm (genetic information of the cow, how many times the cow has given birth, etc.)
- Other derived variables from the above information (e.g., how many times a cow has been diagnosed before, if a cow was treated before, mean, and standard deviation of the change in milk composition in the last 7 and 15 days)

The final delivery of the data for the study conducted by Pakrashi et al. (2023) was in the form of CSV[18] files sent *via* email. The general cow features were included in one CSV file, and milk composition data in another CSV file. Each of the CSV files required special attention for data cleaning, and then, they were joined to make the dataset required. In addition, for the specific task in hand, several derived variables were created for the project, which was not a part of the provided data. This was, then, analyzed, and a machine learning model was trained. While the development was performed, a new batch of data from a subsequent year was available and transferred through email in a similar set of CSV files, which was again combined after fixing a few issues due to incompatibilities with the previous data received.

The above process, if considered in isolation, is relatively straightforward. However, deploying the pilot project in a practical real-world scenario presents a number of problems as follows:

1. As farms generate new data, accessing data *via* email or through a single central data repository is cumbersome.

[18] Comma Separated Value file format.

2. Different teams working on the data coming from the same source had different variable names assigned by the corresponding research teams. There was no clear data description, therefore it was hard to communicate between the teams about the problems.
3. The same data cleaning and transformation were performed independently by the different teams using the same data, leading to redundancy and duplication of effort.
4. The final training dataset prepared by Pakrashi et al. (2023) was found to be useful to other teams. However, processed datasets were sent as a zipped set of CSV files *via* email. This data generation and processing consume a significant amount of time. Additionally, when data are sent *via* email, any updates or changes in processing or data require resending the files, which are often overlooked, leading to continued use of outdated data by other teams.
5. Different teams worked on predicting mastitis through different approaches (Jin et al., 2023). The main target is to predict mastitis in cows using different data available from a dairy farm. Therefore, these projects and the predictive models can be combined to complement each other to build a better mastitis predictor. However, without a shared framework, it is extremely difficult to integrate the products coming from different frameworks.
6. The exact details around data ownership were not clearly defined.

The above issues show that, although valuable data were available from a central source, the processing and analysis work were scattered across silo-ed teams, and the products the teams built (mastitis predictor) were also bound within the team. Aligning with the major challenges defined by WEF (see Section 2), point 1 is a fragmentation issue, while points 2, 3, 5, and 6 relate to standardization. The problems mentioned in points 1, 4, and 5 represent access issues.

4.1.2. Mastitis predictor as a data product in CowMesh

Our proposed data mesh architecture helps address these issues. To demonstrate this, we describe how a collection of projects related to mastitis research can be integrated into the data mesh in an Irish context.

First, the required data mentioned previously (e.g., daily milk data and milk composition data) are aligned with an ontology defined by the semantic layer. This is performed with the help of the data product team and the KS. Such an ontology is shown in Figure 2.

Figure 3 shows how the set of mastitis-related products could be incorporated into the CowMesh. The data domain is named "SCM1_Data" (SCM stands for sub-clinical mastitis), and the six data products are named "SCM1Pred," "SCM1DataAccess," "SCM2Pred," "SCM2DataProcess," "CM1Pred" (CM stands for clinical mastitis), and "HealthMonitor." The roles of these products are explained below.

The "SCM1Pred" product is developed by a research group which is focused on predicting sub-clinical mastitis from daily milking information and historical data about milk and cows. This data product consumes data through protocols set by the semantic layer, by using the API, following the ontology. Therefore, given the ontology (such as Figure 2), this domain will be mainly working with the concepts, such as "Cow," "Milk," and "Disease." The outputs of this product are the predictions and the trained machine learning model. Such a product sharing the outputs of the predictions can be incorporated into a report such as the "CowReport" data product or directly sent to the farmer to assist in decision-making. In addition, potentially the trained machine learning model may be required by other research teams so that they can build another product on top of this (e.g., explaining the predicted outcome). Some additional variables were developed in this project, which were found to be useful to other teams. Therefore, the data product "SCM1DataAccess" provides an API through which the additional variables are accessible.

The data product "CM1Pred" is developed by a research team which works on predicting clinical mastitis. This product accesses data through the semantic layer and uses the additional variables through "SCM1DataAccess," as well as the predictions from "SCM_Pred" as required. Here, "CM1Pred" does not have to recompute the additional variables or receive the variables through cumbersome CSV files.

On the other hand, "SCM2Pred" is a data product which performs sub-clinical mastitis prediction but uses a different perspective and a different set of variables. In this project, some of the data, "SCM2Pred," which are owned privately and are not allowed to be kept. The data pre-processing logic is provided as a service through the "SCM2DataProcess" data product. Therefore, this product can be accessed to use the same data pre-processing while accessing the actual data through the semantic layer from the same sources. The access and sharing of the data, which might be privately held, are governed by the semantic layer. The sharing protocol and other agreements would have been already done while integrating the related data domains.

The data product "HealthMonitor" can be a dashboard summarizing the cows' health in the farms. This takes data from the semantic layer, summarizing the outputs of "CM1Pred," "SCM1Pred," and "SCM2Pred."

The results of these products are also consumed by the report generation product of the semantic layer data domain, such as "CowReport." Including such information in the report can provide dairy farmers and stakeholders with a high-level picture of the health of large herds. Additionally, incorporating mastitis prediction information into the report can prevent farm losses, address cattle welfare, and improve milk production by enabling farmers to take preventive measures ahead of time, before a cow shows the signs of clinical or sub-clinical mastitis.

Issues 1, 2, and 3 are addressed by the ontology and the uniform API provided by the semantic layer and central cleaned accessible data kept with the semantic layer. Issues 4 and 5 are addressed by the data products "SCM1DataAccess" and "SCM1Pred," respectively, which enable the use of the data sharing. These data products, like any other data products in the mesh, would also adhere to the semantic layer ontology.

As each of the data domains is responsible for managing their data by adhering to the semantic layer ontology, cleaning, pre-processing, generating derived attributes, and keeping them updated, the data quality issue is addressed. The access of the data is limited through the API and governed by the semantic layer. Therefore, access can be controlled and the chain of ownership of the data can be tracked, resulting in better data transparency and clear ownership boundaries, thus also addressing issue 6.

To be a part of the CowMesh, it is necessary to interact with the KS to ensure compliance with the protocols. The KS, working with the corresponding products, will assist the relevant team to become integrated into the CowMesh.

4.2. MilkMap

Ireland has a seasonal milk supply influenced by changing weather, soil nutrition, and a host of other environmental and societal factors. In addition to milk yield, the primary change in milk across a season is its composition, i.e., variation in macro (protein, fat, and lactose) and micro (minerals, vitamins, and other bio-actives) constituents. Milk composition determines the yield of dairy products produced by a farm or processor, which can be logistically complex given the fractionation, fermentation, and preservation techniques, and end applications being employed across the sector. Moreover, the compositional makeup of milk determines its functionality, processability, and ultimately its final end use (i.e., as a consumer food or ingredient in another food). For example, consistent manufacture of milk gel-based products such as cheese, casein, or yogurt is highly dependent on the protein and mineral composition of milk. Another example is the relationship between nutrient composition (i.e., protein, minerals, and other ionic species) and heat stability of milk.

4.2.1. MilkMap: traditional process

The MilkMap system is designed to visually represent dairy processing in Ireland. This specialized tool necessitates significant custom processing of the aforementioned raw dairy values. This mapping application is designed to provide additional analytical capabilities in an agricultural context, including the monitoring of dairy production and the provision of time series forecasting (e.g., for yield and composition). While the architecture employed in the MilkMap system considers generalized geographical patterns over long period of time, it also provides the means to drill down into specific localized regions to explore the trends and patterns specific to each region. The final mapping application is ultimately made available to stakeholders across the organization, allowing them to leverage the full potential of the data originally collected. The processed data used in the mapping project are also available in a form that can easily be employed in different systems within the organization.

The delivery of this specialized application for dairy processing in Ireland necessitates significant custom processing of the aforementioned raw dairy values. For instance, the anonymization of dairy data was achieved through transformation of latitude and longitude coordinates to lower resolution H3 hexagonal (Brodsky, 2018) values. These data were transferred from dairy processors to a central repository, allowing for a distributed utilization of this information in application development. No sensitive information is available to either the MilkMap application or the data processing of dairy values needed for it. Other data sources could also be potentially integrated into the MilkMap system. For instance, mid-infra-red (MIR) data taken from existing instruments located across processing plants in Ireland, and grass growth data retrieved from a source such as PastureBase (Hanrahan et al., 2017).

During the development of the MilkMap, several challenges were encountered as follows:

1. Obtaining access to the data was a major hurdle, particularly from the dairy cooperatives, due to legal and privacy concerns. Extensive negotiations and agreements were required to obtain access to the information necessary for the successful implementation of the MilkMap.
2. Combining data from multiple sources was complex and time-consuming, as each source presented its own obstacles to access. Updating and accessing the data were also problematic and necessitated a special Secure File Transfer Protocol (SFTP) connection set up in each case.
3. Noisy data, missing values, and a lack of standardization in data formats presented additional issues, as did the inability to directly communicate with dairy farms or cooperatives, with interaction predominantly passing through the ICBF. Outlier detection was necessary to handle anomalous values, however, pinpointing their root cause proved difficult due to the involvement of numerous intermediaries in the data sharing pipeline.
4. Privacy concerns limited the specificity of the data, making it infeasible to drill down to the level of individual farms when performing detailed analysis.

4.2.2. MilkMap as a data product in CowMesh

Overcoming these challenges required careful consideration of issues around data cleaning, standardization, governance, and security measures to ensure the accuracy and completeness of the data. To reduce the overhead in terms of time and effort, we propose that the MilkMap system could be implemented as a data product within the data mesh, which would help mitigate the issues listed above. The specific benefits of this approach are as follows:

1. To clarify data access issues, such as those around privacy and legal requirements, the semantic layer would prove useful. Since CowMesh provides increased trust and transparency around data usage, we believe that the data sources are more likely to join the mesh and supply the required data.
2. The issue of cumbersome processing of data coming from different sources with different access protocols and data structures can be addressed by the uniform data ontology and data access protocols set by the semantic layer.
3. The issue around noisy and inconsistent data can be handled at the semantic layer when integrating the data domain into the CowMesh. By conforming to the semantic layer protocols, consistent practices for data cleaning, and standardization can be enforced. In addition, since the data domain must now adhere to the data ontology, each feature in the data will be documented (e.g., in terms of range and relationship to other features).

4. The privacy and anonymization issues can be handled at the semantic layer or at the corresponding data domain. In either case, the data handling pipeline can be organized such that data products only have access to a subset of data that is necessary and sufficient as controlled by the data domain.

5. Advantages and opportunities for CowMesh

The objective of the CowMesh is to add value to the dairy industry through intelligent data processing. To do this, the products need access to clean data from the farms to act upon. CowMesh provides the following main advantages which encourage farms and data sources to share data, such that products can provide analytical and predictive insights, which adds value to the dairy farms.

- **Trust**: Data domains and data products can rely on the accuracy and quality of the data they are working with, as well as the security of the data and the trustworthiness of the data sources. By establishing trust in the data and the data sources, organizations can make better decisions and extract more value from the data.
- **Privacy and transparency**: By using the decentralized property of the data mesh, data can be privately held. The nature of the CowMesh allows the private data domains to decide how much data they want to make accessible to the CowMesh. In addition, the chain of ownership can be tracked and audited, ensuring transparency and compliance.
- **FAIR data**: FAIR refers to **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. By ensuring that data follow the data ontology and the access is standardized through the set of access protocols and interfaces (e.g., well-defined APIs), organizations can make it easier for users to discover and access relevant data, as well as combine and analyze data from multiple sources. This can lead to more accurate insights and better decision-making.
- **Decentralized interoperability**: CowMesh combines the contrasting characteristics of data mesh and data fabric and enables data and products to be decentralized while providing one central protocol to be followed, therefore enabling interoperability.
- **Governance**: Governance becomes easier, as the CowMesh follows the standards and protocols set by the semantic layer. This makes the different products follow the standards and protocols easily.

These points address the three main challenges (fragmentation, standards, and access) defined by WEF in the agricultural domain as mentioned in Section 2. Trust, Privacy and Transparency, and Governance can bring cultural changes with respect to how data are shared, which can encourage more organizations and individuals to share their data, addressing the challenge of "standards," whereas FAIR Data and Decentralized Interoperability address the challenges of "fragmentation" and "access" issues.

Several resources and infrastructures need to be maintained to run the CowMesh including designing the specific architecture, hiring a KS, maintaining the APIs and the semantic layer, and maintaining security of the systems and servers. This would require some funds to be spent on CowMesh. As CowMesh is a service which the data domains and data products will use; several aspects of the CowMesh can be monetized, which can help maintain the framework. In addition, through monetization, the individual owners of data (e.g., farmers) and the larger organizations can benefit by charging for the data in an on-demand fashion. This monetary incentive may help more data sources to contribute to CowMesh and get compensated for their data while keeping the data ownership to the corresponding source and maintaining transparency. Some of the ways in which CowMesh can be monetized are as follows:

- **Data use**: The data are accessed through the API, and the use is tracked by the semantic layer. Using this, the total data used by each data product can be tracked and charged. This charge can, then, be distributed to the farmers and the institutions generating the data.
- **Data access**: The uniform data access, interoperability, and also possibly cleaned data are provided by the semantic layer of the CowMesh. Therefore, this service provided by the semantic layer can be monetized. For example, the access to the API calls (not the data) as well as the integration to the CowMesh can be monetized.
- **Report or dashboard consultation**: The semantic layer can generate a periodic report data insights and the analytical components of the CowMesh, for which the organizations and the farmers can pay. This can be a direct value added to the farming industry to see the higher level picture and enable the farmers and the organizations to take updated and informed decisions.
- **Predictive analytics**: Advanced products, such as predicting mastitis, lameness, or ketosis in cows, can be treated as add-on services, which can be offered to farmers on a subscription basis. This is a direct benefit to the farm, as these predictive data products consume data from the farms and then feed back predictive insights to help the farmers.

While the directions above indicate the potential opportunities that the CowMesh can bring, a detailed analysis of the monetization of the CowMesh is outside the scope of the current study.

The different consumers of CowMesh include farmers, researchers, commercial dairy organizations, and veterinary institutions. Farmers can contribute data to CowMesh through data products or the semantic layer. Research organizations, veterinary institutions, and commercial dairy organizations execute specific projects aligned with dairy industry needs and farmers' requirements. Project results benefit farmers through reports and other data products. Research findings can also be shared with commercial dairy organizations for commercialization or as reports. Each party can seamlessly integrate *via* CowMesh, retaining data ownership and transparency of usage, while benefiting from services and products offered, contributing to the improvement of the dairy industry at a national level.

6. Discussion

This study presents a data architecture, CowMesh, designed to unify disparate dairy industry data under a uniform, interoperable, and decentralized framework, thus enabling the products using the data to create value for the dairy farmers and the dairy industry. CowMesh is a combination of data mesh and data fabric. The data mesh's functionality helps the different data products using the dairy data to operate in a manner that is independent and decentralized. The central data product semantic layer is a data fabric, which provides a single unified data model and protocol. This enables connection to and integration of data from different sources within the data mesh. This enables uniform governance, creates trust, promotes data transparency, and keeps the data and the data products decentralized while providing interoperability within the data products. In future, a similar framework can be developed for other agricultural industries tailored for their specific requirements. In addition, specific details about the governance and semantic layer can be explored in an Irish context. Finally, a pilot project to implement CowMesh should be explored in future.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

CG conceptualized the architecture. AP and BM enhanced the architecture and specific cases. AP, CG, and BM contributed to conception and the design of the architecture and designed the mastitis prediction case study. DW and DG designed and wrote the MilkMap case study. AP wrote and finalized the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and the Department of Agriculture, Food, and Marine on behalf of the Government of Ireland under grant numbers [16/RC/3835] and [12/RC/2289_P2].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Albert, P., Saadeldin, M., Narayanan, B., Fernandez, J., Namee, B. M., Hennessey, D., et al. (2022). "Unsupervised domain adaptation and super resolution on drone images for autonomous dry herbage biomass estimation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (New Orleans, LA), 1635–1645. doi: 10.1109/CVPRW56347.2022.00170

Altay, Y., and Albayrak Delialioğlu, R. (2022). Diagnosing lameness with the random forest classification algorithm using thermal cameras and digital colour parameters. *Mediterr. Agric. Sci.* 35, 47–54. doi: 10.29136/mediterranean.1065527

Anglart, D., Hallén-Sandgren, C., Emanuelson, U., and Rönnegård, L. (2020). Comparison of methods for predicting cow composite somatic cell counts. *J. Dairy Sci.* 103, 8433–8442. doi: 10.3168/jds.2020-18320

Barkema, H., von Keyserlingk, M., Kastelic, J., Lam, T., Luby, C., Roy, J.-P., et al. (2015). Invited review: changes in the dairy industry affecting dairy cattle health and welfare. *J. Dairy Sci.* 98, 7426–7445. doi: 10.3168/jds.2015-9377

Bauer, E. A., and Jagusiak, W. (2022). The use of multilayer perceptron artificial neural networks to detect dairy cows at risk of ketosis. *Animals* 12, 332. doi: 10.3390/ani12030332

Bhat, R., Di Pasquale, J., Bnkuti, F. I., Siqueira, T. T. d. S., Shine, P., and Murphy, M. D. (2022). Global dairy sector: trends, prospects, and challenges. *Sustainability* 14, 4193. doi: 10.3390/su14074193

Bobbo, T., Biffani, S., Taccioli, C., Penasa, M., and Cassandro, M. (2021). Comparison of machine learning methods to predict udder health status based on somatic cell counts in dairy cows. *Sci. Rep.* 11, 1–10. doi: 10.1038/s41598-021-93056-4

Bode, J., Kühl, N., Kreuzberger, D., and Hirschl, S. (2023). Data mesh: motivational factors, challenges, and best practices. *arXiv preprint arXiv:2302.01713.* doi: 10.48550/arXiv.2302.01713

Borchers, M., and Bewley, J. (2015). An assessment of producer precision dairy farming technology use, prepurchase considerations, and usefulness. *J. Dairy Sci.* 98, 4198–4205. doi: 10.3168/jds.2014-8963

Brodsky, I. (2018). *H3: Hexagonal Hierarchical Geospatial Indexing System.* Uber Open Source.

Butte, V. K., and Butte, S. (2022). "Enterprise data strategy: a decentralized data mesh approach," in *2022 International Conference on Data Analytics for Business and Industry (ICDABI)* (Sakhir), 62–66. doi: 10.1109/ICDABI56818.2022.10041672

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., et al. (2013). The agrovoc linked dataset. *Seman. Web* 4, 341–348. doi: 10.3233/SW-130106

Cavero, D., Tölle, K.-H., Rave, G., Buxadé, C., and Krieter, J. (2007). Analysing serial data for mastitis detection by means of local regression. *Livestock Sci.* 110, 101–110. doi: 10.1016/j.livsci.2006.10.006

Chianese, D., Rotz, C. A., and Richard, T. L. (2009). Simulation of nitrous oxide emissions from dairy farms to assess greenhouse gas reduction strategies. *Trans. ASABE* 52, 1325–1335. doi: 10.13031/2013.27782

Christ, J., Visengeriyeva, L., and Harrer, S. (2022). *Data Mesh Architecture - Data Mesh From an Engineering Perspective*. Available online at: https://www.datamesh-architecture.com (accessed February 24, 2023).

Dehghani, Z. (2022a). *Data Mesh*. Marcombo.

Dehghani, Z. (2022b). *Data Mesh Principles and Logical Architecture*. Available online at: https://martinfowler.com/articles/data-mesh-principles.html (accessed August 17, 2023).

Dolhopolov, A., Castelltort, A., and Laurent, A. (2023). *Implementing a Blockchain-Powered Metadata Catalog in Data Mesh Architecture*. Available online at: https://hal.umontpellier.fr/hal-04156134

Douphrate, D. I., Hagevoort, G. R., Nonnenmann, M. W., Lunner Kolstrup, C., Reynolds, S. J., Jakob, M., et al. (2013). The dairy industry: a brief description of production practices, trends, and farm characteristics around the world. *J. Agromed*. 18, 187–197. doi: 10.1080/1059924X.2013.796901

Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., and Petrovski, K. R. (2019). Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: deep learning and gradient-boosted trees outperform other models. *Comput. Biol. Med*. 114, 103456. doi: 10.1016/j.compbiomed.2019.103456

Fadul-Pacheco, L., Wangen, S. R., da Silva, T. E., and Cabrera, V. E. (2022). Addressing data bottlenecks in the dairy farm industry. *Animals* 12, 721. doi: 10.3390/ani12060721

Fletcher, G., Groth, P., and Sequeda, J. (2020). Knowledge scientists: unlocking the data-driven organization. *arXiv preprint arXiv:2004.07917*. doi: 10.48550/arXiv.2004.07917

Gabriel, A., and Gandorfer, M. (2023). Adoption of digital technologies in agriculture-an inventory in a european small-scale farming region. *Precis. Agric*. 24, 68–91. doi: 10.1007/s11119-022-09931-1

Gargiulo, J., Eastwood, C., Garcia, S., and Lyons, N. (2018). Dairy farmers with larger herd sizes adopt more precision dairy technologies. *J. Dairy Sci*. 101, 5466–5473. doi: 10.3168/jds.2017-13324

Goedegebuure, A., Kumara, I., Driessen, S., Di Nucci, D., Monsieur, G., Heuvel, W.-J. V. D., et al. (2023). Data mesh: a systematic gray literature review. *arXiv preprint arXiv:2304.01062*. doi: 10.48550/arXiv.2304.01062

Halasa, T., Huijps, K., Østerås, O., and Hogeveen, H. (2007). Economic effects of bovine mastitis and mastitis management: a review. *Vet. Q*. 29, 18–31. doi: 10.1080/01652176.2007.9695224

Hanrahan, L., Geoghegan, A., O'Donovan, M., Griffith, V., Ruelle, E., Wallace, M., et al. (2017). PastureBase Ireland: a grassland decision support system and national database. *Comput. Electron. Agric*. 136, 193–201. doi: 10.1016/j.compag.2017.01.029

Hansen, B. G., Herje, H. O., and Höva, J. (2019). Profitability on dairy farms with automatic milking systems compared to farms with conventional milking systems. *Int. Food Agribus. Manage. Rev*. 22, 215–228. doi: 10.22434/IFAMR2018.0028

Heringstad, B., Klemetsdal, G., and Ruane, J. (2000). Selection for mastitis resistance in dairy cattle: a review with focus on the situation in the Nordic countries. *Livestock Prod. Sci*. 64, 95–106. doi: 10.1016/S0301-6226(99)00128-1

Hooshmand, Y., Resch, J., Wischnewski, P., and Patil, P. (2022). From a monolithic PLM landscape to a federated domain and data mesh. *Proc. Des. Soc*. 2, 713–722. doi: 10.1017/pds.2022.73

Jakku, E., Taylor, B., Fleming, A., Mason, C., Fielke, S., Sounness, C., et al. (2019). "If they don't tell us what they do with it, why would we trust them?" Trust, transparency and benefit-sharing in smart farming. *Wageningen J. Life Sci*. 90–91, 1–13. doi: 10.1016/j.njas.2018.11.002

Jin, C., Upton, J., and Namee, B. M. (2023). "Do cow's have fingerprints? Using time series techniques and milk flow profiles to characterise cow behaviours and detect health issues," in *8th Workshop on Advanced Analytics and Learning on Temporal Data (AALTD 2023)* (Torino).

Joshi, D., Pratik, S., and Rao, M. P. (2021). "Data governance in data mesh infrastructures: the Saxo bank case study" in *ICEB 2021 Proceedings* (Nanjing, China), 52. Available online at: https://aisel.aisnet.org/iceb2021/52

Kadam, P., and Vijayumar, S. (2018). "Prediction model: CO2 emission using machine learning," in *2018 3rd International Conference for Convergence in Technology (I2CT)* (Pune), 1–3. doi: 10.1109/I2CT.2018.8529498

Larmande, P., and Todorov, K. (2021). "Agrold: a knowledge graph for the plant sciences," in *The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, Virtual Event, Proceedings, Vol. 12922 of Lecture Notes in Computer Science*, eds A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. M. Barnaghi, A. Haller, M. Dragoni, and H. Alani (Cham: Springer), 496–510. doi: 10.1007/978-3-030-88361-4_29

Martin, N., Russelle, M., Powell, J., Sniffen, C., Smith, S., Tricarico, J., et al. (2017). Invited review: sustainable forage and grain crop production for the US dairy industry. *J. Dairy Sci*. 100, 9479–9494. doi: 10.3168/jds.2017-13080

O'Hara, R., Zimmermann, J., and Green, S. (2021). A multimodality test outperforms three machine learning classifiers for identifying and mapping paddocks using time series satellite imagery. *Geocarto Int*. 37:9748–9766. doi: 10.1080/10106049.2021.2024278

Pakrashi, A., Ryan, C., Gueret, C., Berry, D., Corcoran, M., Keane, M. T., et al. (2023). Early detection of subclinical mastitis in lactating dairy cows using cow level features. *J. Dairy Sci*. 106, 4978–4990. doi: 10.3168/jds.2022-22803

Petrovski, K., Trajcev, M., and Buneski, G. (2006). A review of the factors affecting the costs of bovine mastitis. *J. South Afr. Vet. Assoc*. 77, 52–60. doi: 10.4102/jsava.v77i2.344

Pongpech, W. A. (2023). A distributed data mesh paradigm for an event-based smart communities monitoring product. *Proc. Comput. Sci*. 220, 584–591. doi: 10.1016/j.procs.2023.03.074

Ryan, C., Guéret, C., Berry, D., Corcoran, M., Keane, M. T., and Mac Namee, B. (2021). Predicting illness for a sustainable dairy agriculture: predicting and explaining the onset of mastitis in dairy cows. *arXiv preprint arXiv:2101.02188*. doi: 10.48550/arXiv.2101.02188

Shahinfar, S., Khansefid, M., Haile-Mariam, M., and Pryce, J. (2021). Machine learning approaches for the prediction of lameness in dairy cows. *Animal* 15, 100391. doi: 10.1016/j.animal.2021.100391

Viguier, C., Arora, S., Gilmartin, N., Welbeck, K., and O'Kennedy, R. (2009). Mastitis detection: current trends and future perspectives. *Trends Biotechnol*. 27, 486–493. doi: 10.1016/j.tibtech.2009.05.004

Wang, H., Guo, T., Wang, Z., Xiao, J., Gao, L., Gao, X., et al. (2023). Precowketosis: a shiny web application for predicting the risk of ketosis in dairy cows using prenatal indicators. *Comput. Electron. Agric*. 206, 107697. doi: 10.1016/j.compag.2023.107697

WEF (2023). *Here's How the Agricultural Sector Can Solve Its Data Problem*. World Economic Forum. Available online at: https://www.weforum.org/agenda/2023/01/here-s-how-agricultural-sector-data-problem-davos2023/ (accessed August 17, 2023).

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi: 10.1038/sdata.2016.18

Wolfert, S., Ge, L., Verdouw, C., and Bogaardt, M.-J. (2017). Big data in smart farming-a review. *Agric. Syst*. 153, 69–80. doi: 10.1016/j.agsy.2017.01.023

Wysel, M., Baker, D., and Billingsley, W. (2021). Data sharing platforms: how value is created from agricultural data. *Agric. Syst*. 193, 103241. doi: 10.1016/j.agsy.2021.103241

Yalcin, C., Stott, A., Logue, D., and Gunn, J. (1999). The economic impact of mastitis-control procedures used in Scottish dairy herds with high bulk-tank somatic-cell counts. *Prevent. Vet. Med*. 41, 135–149. doi: 10.1016/S0167-5877(99)00052-5

Zhou, X., Xu, C., Wang, H., Xu, W., Zhao, Z., Chen, M., et al. (2022). The early prediction of common disorders in dairy cows monitored by automatic systems with machine learning algorithms. *Animals* 12, 1251. doi: 10.3390/ani12101251

frontiers | Frontiers in Artificial Intelligence

TYPE Methods
PUBLISHED 16 October 2023
DOI 10.3389/frai.2023.1187090

OPEN ACCESS

EDITED BY
Matthew McCabe,
King Abdullah University of Science and Technology, Saudi Arabia

REVIEWED BY
Shailesh Tripathi,
Tampere University of Technology, Finland
Damion M. Dooley,
Simon Fraser University, Canada

*CORRESPONDENCE
Baptiste Darnala
✉ baptiste.darnala@elzeard.co
Clément Jonquet
✉ clement.jonquet@inrae.fr

RECEIVED 15 March 2023
ACCEPTED 28 August 2023
PUBLISHED 16 October 2023

CITATION
Darnala B, Amardeilh F, Roussey C, Todorov K and Jonquet C (2023) C3PO: a crop planning and production process ontology and knowledge graph. *Front. Artif. Intell.* 6:1187090. doi: 10.3389/frai.2023.1187090

C3PO: a crop planning and production process ontology and knowledge graph

Baptiste Darnala[1,2]*, Florence Amardeilh[2], Catherine Roussey[3], Konstantin Todorov[1] and Clément Jonquet[1,3]*

[1]LIRMM, CNRS, University of Montpellier, Montpellier, France, [2]Elzeard, Bordeaux, France, [3]MISTEA, INRAE, Institut Agro, University of Montpellier, Montpellier, France

Vegetable crop farmers diversify their production by growing a range of crops during the season on the same plot. Crop diversification and rotation enables farmers to increase their income and crop yields while enhancing their farm sustainability against climatic events and pest attacks. Farmers must plan their agricultural work per year and over successive years. Planning decisions are made on the basis of their experience regarding previous plans. For the purpose of assisting farmers in planning decisions and monitoring, we developed the *Crop Planning and Production Process Ontology* (C3PO), i.e., a representation of agricultural knowledge and data for diversified crop production. C3PO is composed of eight modules to capture all crop production dimensions and complexity for representing farming practices and constraints. It encodes agricultural processes and farm plot organization and captures common agricultural knowledge. C3PO introduces a representation of technical itineraries, i.e., sequences of technical farming tasks to grow vegetables, from soil identification and seed selection to harvest and storage. C3PO is the backbone of a knowledge graph which aggregates data from heterogeneous related semantic resources, e.g., organism taxonomies, chemicals, reference crop listings, or development stages. C3PO and its knowledge graph are used by the Elzeard enterprise to develop knowledge-based decision support systems for farmers. This article describes how we built C3PO and its knowledge graph—which are both publicly available—and briefly outlines their applications.

KEYWORDS

ontology, semantic resource, knowledge representation, knowledge graph, agriculture, crop planning, crop production process, digital farming

1. Introduction

Agricultural work is complex, farmers need to take various factors such as weather, seasonality, commercial demand, and plant life cycles into account when planning production. Moreover, consumers' and farmers' behavior and practices are changing, with greater consideration for ecological and economic aspects. After the WWII, single-crop farming and the use of chemical inputs were highly promoted, but this led to reduced soil fertility and chemical contamination of soil and crops. In recent years, farmers have begun adopting agroecological practices. Agroecology offers a way of designing farming production systems that rely on agroecosystem functionalities. Crop diversity is a key aspect of this approach. Several scientific studies (Isbell et al., 2017; Paut et al., 2019) have shown that, in both spatial and temporal terms, crop plant diversity (i) improves risk management under changing weather and economic conditions; (ii) delivers a natural defense system against diseases and pest attacks; and (iii) increases agroecosystem stability and resilience. Farmers must consider different stages of development, treatments, and biological interactions between plants. However, these new agroecological practices increase farmers' workload, management complexity, and mental burden (Morel and Léger, 2015; Dumont, 2021). Between 2019 and 2020, the SME Elzeard conducted over 150 interviews with vegetable farmers, agricultural advisors, teachers, and researchers.[1] Elzeard identified several technological barriers, including the need for knowledge sharing and new operational tools to assist vegetable farmers in their daily crop management, i.e., optimizing crop rotations, yields, and finding alternatives to chemical inputs.

Agricultural knowledge–i.e., consensual information used by farmers to make decisions and to take actions–is currently scattered webwide, in books, archives, and databases while also conveyed informally through interpersonal communication and cultural practices. As well, climate change is increasing an important factor to mitigate in agriculture. Climate has a direct impact, with rainfall fluctuations and heat waves, affecting plant growth and promoting disease emergence (Mendelsohn, 2009; Arora, 2019). To help farmers face this new challenge and its impact on agriculture, it is necessary to share and discover new knowledge. Climate change and the adoption of agroecology approaches now call for the development of novel knowledge-based systems. Knowledge needs to be formalized and shared with anyone who needs it, requiring a common formalization. Moreover, quality knowledge must be readily findable, accessible, interoperable, and reusable (FAIR) for users (Wilkinson et al., 2016). In that respect, semantic web technologies facilitate the development of robust knowledge-based systems by enabling formalization and sharing of knowledge for validation, enrichment, and discovery with native reasoning. Common formalization also facilitates the aggregation of diverse resources and provenance verification to control the quality of shared knowledge.

Semantic Web technologies are used to build knowledge graphs (KG) and ontologies in agriculture (crops and livestock farming, etc.) for both experimental and agricultural purposes (Drury et al., 2019). For instance, AgroPortal (Jonquet et al., 2018) hosts approximately 150 ontologies and vocabularies (as of January 2023), many of which focus on agrifood-environment issues, such as the Agronomy ontology (Devare et al., 2016) and FoodOn (Dooley et al., 2018). However, diversified vegetable farming has not been extensively explored, and there is a substantial need for research in this area. Diversified vegetable crop farmers define technical itineraries (*"itinéraire cultural"* in French, abbreviated in the domain as ITK) as sequences of technical farming tasks to grow vegetables from soil identification and seed selection to harvest and storage. In these sequences, every task and their timing depend on each other and other parameters such as the cropping mode (open field or under cover) or the climatic conditions. As an illustration, planting an onion crop in winter can lead to a spring harvest if cultivated under cover. However, if the onions are cropped in an open field, they will be harvested in the summer. Farmers and agricultural experts define the technical itineraries and draw up plans for the following year(s) based on what happened previously in the fields. They can share their technical itineraries with other farmers who will adapt them to their specific agricultural context, as defined by parameters such as soil type and climate, the number of farm workers or the diversity of crops. To the best of our knowledge, technical itineraries have not been represented using Semantic Web technologies.

In this study, we developed the *Crop Planning and Production Process Ontology* (C3PO) and populated it into a KG. The couple jointly captures vegetable crop farming management concepts and knowledge to support multiple applications in diversified crop planning. The ontology incorporates the representation of technical itineraries for farm planning and management, as well as the representation of plants, plot organization, and chemical products and equipment. The KG is aligned with other agricultural Semantic Web resources to integrate reference data dispersed in organization systems. Our goal is to represent interactions between living organisms, farmers' actions, and input products. Each type of entity is identified and managed under different web standards. This study presents the methodology by which the ontology and related KG have been developed. A subpart of the C3PO knowledge graph will soon be publicly available on *La Serre des Savoirs*, a web portal that pools integrated and harmonized knowledge about plants and farming practices. In addition to these knowledge assets, we are building multiple applications such as two knowledge-based decision support systems to assist farmers: *Elzeard*, a web and mobile application to plan and monitor crop production in vegetable farming systems; and *La Pépinière*, a free web application to help beginners to design their farms and their future production system.

The rest of the study is organized as follows: Section 2 presents related work in ontology development methodology, agriculture, and semantics; Section 3 outlines the methodology implemented to build the ontology and knowledge graph; Section 4 presents the ontology and knowledge graph; Section 5 presents applications based on the ontology and knowledge graph; Section 6 discusses the problems encountered and limitations of our study; and Section 7 concludes and presents the perspectives.

2. Related work

Many ontology and knowledge graph development methodologies have been created to support ontology development since the 1990s. In this section, we present those used in the development of our resource. The NeOn methodology (Suárez-Figueroa et al., 2012) presents nine flexible scenarios to build an ontology and "ontology networks" based on the reuse of semantic resources, the transformation of non-semantic resources, and the reuse of ontology design patterns. An ontology network is a collection of ontologies linked *via* relations such as mapping, import, or version. By this methodology, we designed C3PO in multiple ontology modules to address different scenarios. The list of C3PO's modules is presented in Section 4.1. According to the NeOn methodology definition, an ontology module is "a part of the ontology that defines a relevant set of terms". However, although NeOn offers interesting guidelines to organize the ontology development, the methodology is time-consuming due to the required quantity of documentation and the NeOn toolkit,

[1] https://elzeard.co

i.e., the integrated development environment proposed with the methodology, is no longer updated. The Agile methodology was originally used for the development of systems and applications, then later also for ontologies, which implies iterative development and publication and continuous collaboration with consumers. SAMOD (Peroni, 2016) is an Agile methodology with which ontologists develop small ontology iterations for describing a particular use-case and addressing competency questions. After review, the iteration is added to the main ontology representing the whole domain. We adopted this approach for the development of each C3PO module. The Linked Open Terms (LOT) methodology (Poveda-Villalón et al., 2022) is another Agile methodology describing each ontology development step, from specification to publication. LOT is focused on industrial projects as the aim to be compatible with software development methodologies with iterative steps. Moreover, a set of tools is provided as well as examples of how they may be used in ontology development. We built C3PO by combining the development steps presented in LOT and the iteration development process presented in SAMOD.

We studied models focused on agronomy and agriculture. We queried AgroPortal to identify ontologies and vocabularies to represent plant knowledge, agricultural tasks, and plot organization and identified multiple semantic resources: the French Crop Usage thesaurus (FCU) (Roussey, 2018), a list of cultivated plants organized by agriculture uses in France; the Agroecology Knowledge Management application (GECO, in French), a research information system for the GECO data graph (Soulignac et al., 2019) to design innovate agroecology-oriented crop systems; TAXREF-LD (Michel et al., 2017), a linked data representation of the national repository of fauna and flora of France; the NCBI Taxonomy, a curated classification and nomenclature for organisms; Plant Ontology (Jaiswal et al., 2005), a structured vocabulary and database resource that links plant anatomy, morphology, growth, and development to plant genomics data; Crop Ontology (Arnaud et al., 2012), a vocabulary of observable characteristics of common crops for food and agriculture; and the AgroLD knowledge graph focused on plant biology data (Larmande and Todorov, 2021).[2] Each of these resources is based on a specific viewpoint but cannot be used alone to represent plant knowledge in agriculture. However, we have combined some of them in a coherent integrated knowledge graph that is presented later. Other ontologies and vocabularies available in AgroPortal-but which we did not directly used in our work- to represent agricultural processes include: the Agronomy Ontology (AGRO) (Devare et al., 2016), which represents agronomic experiments by recording precise observations concerning experiments on agricultural plots but is not geared toward agricultural planning and monitoring. The DEMETER Agriculture Information Model (Palma et al., 2022) focuses on smart farming solutions using sensors to monitor crops, which is currently beyond our scope.

Semantic Web technologies may represent processes which could be used to represent tasks in agriculture. The Provenance Ontology (Prov-O) (Lebo et al., 2013) traces the provenance and evolution of activities, interacting with involving agents and entities. Prov-O is an interesting ontology that needs to be specialized to represent a domain, but it does not address all of our needs, especially with respect to temporal aspects required for representing technical itineraries. Otherwise, it is essential to represent theoretical dates, i.e., dates not related to a year (e.g. 04/25), which is not possible. We, thus, opted to use the Time Ontology (Hobbs and Pan, 2006) and extended it to fulfill our needs. However, Prov-O is used to track KG updates, as explained in Section 3.4. We also studied ValueFlows, an ontology that describes economic value flows according to three representation layers (Knowledge, Plan, and Observation).[3] Knowledge represents plan specification to make something: an ordering set of tasks (e.g., a cooking recipe which specify step-by-step the recommended quantity of ingredients to cook); plan represents the planning of these tasks by an agent and the choice made to implement them (e.g., the actual recipe steps with the quantity of ingredients planned to be used); observation represents the plan execution of the tasks (e.g., the actual recipe steps followed by the cook with the quantities of ingredients used). We used these three layers to conceptualize the principles underlying the representation of technical itineraries: plan specification, plan, and plan execution.

To the best of our knowledge, there are currently no ontology available to help diversified vegetable crop farms in planning and managing their farming tasks. Existing ontologies either only represent sub-parts of the problem or focus on other agricultural sectors, such as cereal cropping or livestock farming. However, the representation of technical itineraries for diversified vegetable agriculture and their use for agricultural planning has not been addressed in an adequate and complete ontology.

3. Requirements and methodology for FAIR ontology building and sharing

As mentioned previously, we built C3PO using a combination of LOT and SAMOD. We followed the LOT workflow, consisting of the ontology: (i) requirement specification, (ii) implementation, (iii) publication, and (iv) maintenance. The methodology also includes the knowledge graph construction and maintenance procedures. In the development, we built a component that meets current needs before adding it to the ontology, as recommended by SAMOD. Combining these methodologies brought us: (i) a general process regarding the construction of the ontology, due to LOT; (ii) a process regarding the update of the ontology though iterative development, thanks to SAMOD. We recommend this combination to any ontology development project related to an application development using agile methodology, with multiple viewpoints, multiple subdomains, and that integrate several and heterogeneous data sources. The main actors are the Domain Expert (DE), i.e., who offers the domain knowledge covered by the ontology and the overall vision of the work, and the Ontology Expert (OE), who has expertise in ontology and knowledge representation methods. Figure 1 presents the C3PO implementation steps, people involved, and tools used.

[2] https://www.ncbi.nlm.nih.gov/taxonomy

[3] https://lab.allmende.io/valueflows

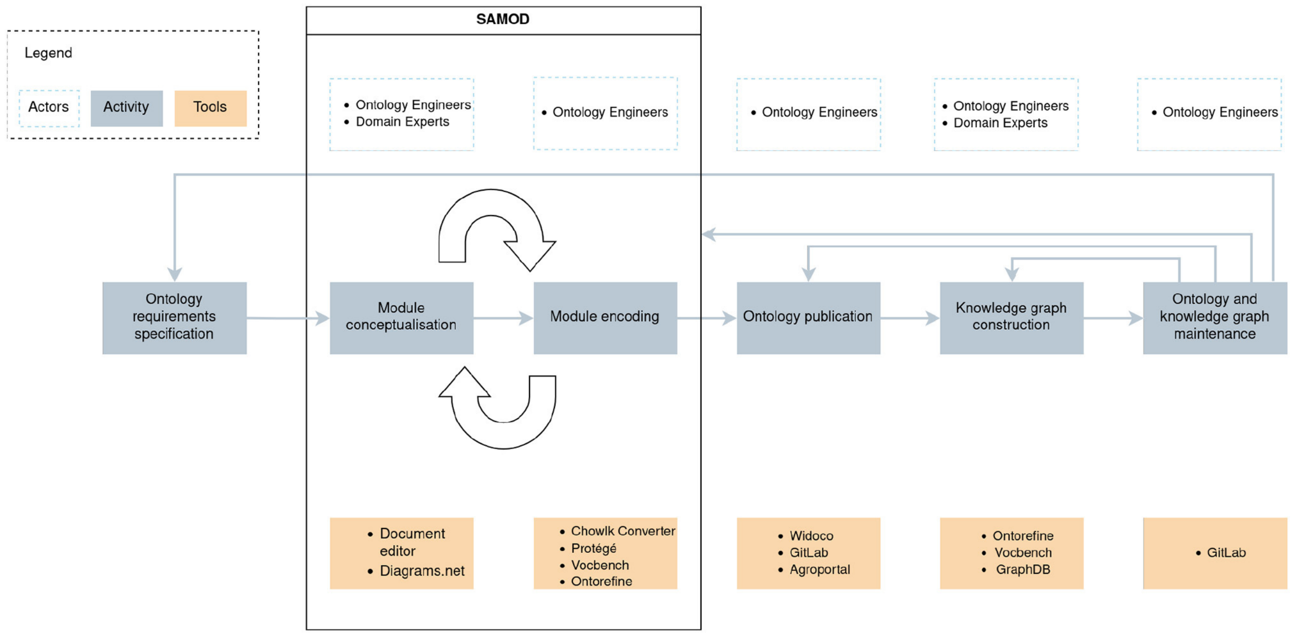


FIGURE 1
C3PO implementation steps in the LOT/SAMOD methodologies and tools used.

3.1. Ontology requirement specification

3.1.1. Use-case specification

This specification involves collecting the requirements for the ontology. The DEs present the needs in terms of data and queries required with informal text and schemas. This study defines the scope of the domain of the ontology should address and specify the use-cases. Moreover, DEs and OEs can extract multiple sub-domains for the ontology, which leads to the creation of several ontology modules. OEs review the document to gain insight into the data need, constraints, and logical inference expected and split the use-case regarding the different domains. Moreover, OEs also distinguish between static qualitative data (e.g., in our case plant families, inputs, crops, physical locations, and laws) and user input data. This difference is important as the implementation needs are not the same: user data will be collected *via* forms or sensors, then curated and eventually analyzed, whereas qualitative data that are relevant for the domain will have to be found by OEs in relevant external knowledge sources in the form of open data, existing ontologies or KGs, community standards, and norms. When such standards do not exist, OEs will have to build them. For instance, the way farmers organize their plots is typically user data, while the taxonomic representation of plants represented in TAXREF-LD is a relevant knowledge source that has been integrated into the C3PO knowledge graph.

3.1.2. Functional ontological requirements

OEs describe each use-case with a title, an informal description, and competency questions related to this use-case to support the ontology development process. Examples of use-case and some competency questions for C3PO are presented in Table 1. Some competency questions are also described in the ontology metadata and in the documentation.[4]

3.2. Ontology implementation

3.2.1. Ontology conceptualization

To conceptualize an ontology module that fulfills the requirements, we opted to use the SAMOD methodology because it offers the possibility to build an ontology module for some use-cases before integration in the main ontology. The proposition at the end of the conception phase is a list of concepts, relations, and queries.

OEs analyze the requirements and extract a list of concepts and relations. They propose a name, URI, and definition for each one. As an example of the use-case in Table 1, an extractable concept is a “crop”, however we choose to create the “CultivatedPlant” class with an URI and associated definition: “Vegetal organism cultivated by human beings”. Then, an ontology module is built with the classes and properties proposed with Chowlk notation (Chávez-Feria et al., 2022), i.e., an UML-based notation to build ontology diagrams in Diagrams.net.[5] The diagram is composed of the classes, properties, and an instantiated data example. An example of a diagram produced and related to the example in Table 1 is presented in Figure 2, this is a representation of the onion, its families, and labels. The individuals of the plants and their families are both typed `skos:Concept` and `owl:Class`.

[4] https://www.elzeard.co/ontologies/c3po/

[5] https://www.drawio.com/

TABLE 1 Use-case description of the representation of plants and their families.

Title	Representation of the organization of plants and their groups
Description	Plants are organized in several families (group of plants that share some common characteristics). These families can describe botanical characteristics, or can describe their usage in agriculture or consumption. Representing only the plant and the families is not sufficient as plants have cultivars that could be split in multiple categories (called varietal types) regarding their physical characteristics. For example, onion can be divided into yellow onion and red onion. The need is to get a representation of the whole plant taxonomy to enhance the farmers' knowledge about plant characteristics.
Competency Questions	1. What are the botanical and usage families of a given crop? Botanical family of the onion is amaryllidaceae and usage family is bulb vegetables. 2. Are these two given crops from the same family? Onion and garlic are from the same botanical family. 3. What are the crops in a given botanical or usage family? Onion, garlic, and shallot are in the amaryllidaceae botanical family.

3.2.2. Ontology conceptualization validation

To validate the conceptualization, we present modeling diagrams to DE validation and refinement. Class and property names are validated with DEs to check their existence in the domain.

3.2.3. Ontology encoding

After validation, OEs integrate the diagram in C3PO diagrams to generate a formal representation in OWL using Chowlk Converter.[6] The class and property definition are written on the side in another tabular file so that they can be collaboratively edited by DEs and OEs. We generate the OWL file of the definition with OntoRefine, i.e., a tool to transform a tabular file into an RDF file using a template.[7] We combine the Chowlk output file and definition file in Protégé (Musen, 2015) to consolidate and export a complete OWL file for C3PO module that we will use in our KG, exchanges with external parties, or publish in AgroPortal.

We supplement the C3PO KG with terms from controlled vocabularies used for property values. This "vocabulary part" of our KG contains information such as climate, unit of measure, and irrigation mode. To ease collaboration, it is maintained with VocBench (Stellato et al., 2015), an application that allows us to build ontology and thesaurus.

3.2.4. Ontology evaluation

To evaluate the ontology module: (i) First, to validate the domain representation, we write SPARQL queries matching the competency questions and executed them on the ontology. The competency questions and the SPARQL queries are available.[8] (ii) Second, to validate the structure and syntax, we use OOPS (Poveda-Villalón et al., 2014) to detect any classic ontology pitfalls. (ii) Third, to validate the embedded logic, we run the Pellet (Sirin et al., 2007) and Hermit (Glimm et al., 2014) reasoners to check the consistency of the ontology regarding the domain. (iv) Finally, to validate the usability, we explicitly used the ontology as a data model for an implemented application, which does run and fulfill its requirements (see Section 5 for details).

3.2.5. FAIR ontology publication

The ontology publication phase consists of providing the OWL files and the documentation online according to FAIR principles. Regarding these principles, we published C3PO on GitLab in the form of a set of ontology module files, as presented in Section 4. We produced the documentation of each module using Widoco (Garijo, 2017), a tool that generates an HTML documentation from an OWL file. The publication phase also consists in adding metadata to the ontology to improve its description. We upload C3PO in AgroPortal and declare some metadata using the AgroPortal metadata schema named MOD (Dutta et al., 2015). To evaluate C3PO's fairness level (i.e., to which level our ontology adhere to the FAIR Principles), we used O'FAIRe (Amdouni et al., 2022), an ontology fairness evaluator for semantic resource proposed by AgroPortal. C3PO reaches 59% of fairness. More information about metadata standard is described in Section 4.2.

3.3. Knowledge graph construction

The knowledge graph constructed under C3PO consists of several heterogeneous data sources. The type of source impacted the way we imported data, as described hereafter.

3.3.1. Domain expert data

As DEs are not Semantic Web experts, we support them in populating the C3PO KG with tabular files. Then, we use OntoRefine as was done when building the ontology to produce RDF files. The process is used to import characteristics of the cultivated plants and technical itineraries in the knowledge graph.

3.3.2. Relational database import

Some agricultural databases that we wanted to be in C3PO KG were not available in RDF. We, thus, use tabular formats of the databases to which we apply a preprocessing, e.g., to produce URIs. In the tabular to RDF transformation process, value sets are encoded directly in our vocabulary that is built with SKOS (Miles and Bechhofer, 2009). Then, the files are transformed in OWL format with OntoRefine to produce an RDF dataset to be imported in the C3PO KG. We use this process for the integration of Basagri, a database containing information on agricultural chemical products and their uses with plants distributed by the Lexagri company.[9] The database is updated daily, but it is not freely available. In section 4.1, we present the whole process and how

[6] https://chowlk.linkeddata.es/

[7] https://www.ontotext.com/products/ontotext-refine/

[8] https://gitlab.com/serre-des-savoirs/c3po/-/blob/master/C3PO-Competency_question.xlsx

[9] https://www.lexagri.com/products/basagri/

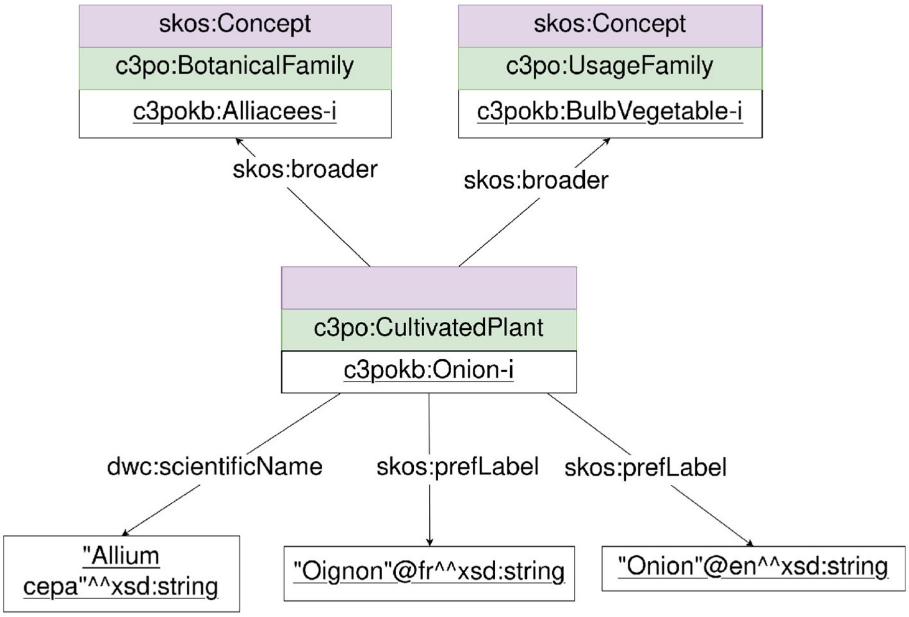


FIGURE 2
Representation of plants and their families.

we used the E-PHY ontology (Bouazzouni and Jonquet, 2021) to represent the Basagri chemical data.

3.3.3. RDF linking

Semantic Web resources exist in agriculture, as presented in the related work. We manually linked our knowledge graph with other KGs such as TAXREF-LD and FCU to enhance the representation of plants with botanical and usage information (Darnala et al., 2022). The process involved DEs to produce and validate the set of links.

3.3.4. Application data

As the ontology is designed to be used for crop planning, part of the data are from user input from applications (*Elzeard*, *La Serre des Savoirs* and *la Pépinière*).

The final C3PO KG is built by importing all the previously revised data in the same RDF database, i.e., GraphDB in our case.[10] GraphDB was chosen to meet our requirements with respect to ease in dealing with RDF data directly, write and test SPARQL queries, create named graphs and reasoning features.

3.4. Ontology and knowledge graph maintenance

The ontology is updated each time a new use-case appears and requires an ontology development. Updates are also done to fix bugs remaining in the ontology or the knowledge graph. Moreover, as the ontology is published on GitLab, the submission of regarding bugs or improvements is possible. To update the knowledge graph, we implemented pipelines to produce RDF graphs from CSV files to enable continuous data development by DEs and improvement of the knowledge graph without an extensive need of an OE. To prevent direct insertion of triples in the knowledge graph and possible errors, we promote building of a new named graph for static data, i.e., in our case plant and input knowledge, each time a new batch of data is imported in the current knowledge graph. However, we could have problems of changing ids between two version of the knowledge graph, so a backup of each version of the knowledge graph is required.

Regarding user data present in the knowledge graph, we use Prov-O to track updates and provenance. Each update lists the modified instances, the user involved, and the time of the update. The update description is saved in JSON format as a value of a data property in the knowledge graph. Tracking updates allows us to know who performs the update and recover from previous timestamps if needed.

4. The C3PO ontology and knowledge graph

In this section, we describe the current version of the C3PO and specific development strategies for both the ontology source file and the knowledge graph. We divided the ontology in several modules, each representing a specific sub-domain of interest for vegetable agriculture planning. We chose such modular representation as we identified/extracted from the conceptualization step several sub-domains, which could work independently but still strongly related. Furthermore, modules helped during the conceptualization

[10] https://graphdb.ontotext.com/

TABLE 2 C3PO modules and their URIs and namespaces.

Module name	Module URI	Module namespace
Support modules		
Time	http://www.elzeard.co/ontologies/c3po/time	c3potime
Vocabulary	http://www.elzeard.co/ontologies/c3po/vocabulary	c3povocab
Parameter	http://www.elzeard.co/ontologies/c3po/parameter	c3poparam
Domain modules		
Plant	http://www.elzeard.co/ontologies/c3po/plant	c3poplant
Plot	http://www.elzeard.co/ontologies/c3po/plot	c3poplot
Crop Management	http://www.elzeard.co/ontologies/c3po/cropManagement	c3pocm
Admin	http://www.elzeard.co/ontologies/c3po/admin	c3poadmin
Supply	http://www.elzeard.co/ontologies/c3po/supply	c3posupply
Sale	http://www.elzeard.co/ontologies/c3po/sale	c3posale

to divide the work and focus on sub-domains instead of the whole area. As an example, we divided the representation of plant knowledge and plot organization into two distinct modules. We divided the modules between support modules and domain modules. Support modules are used in almost all the domain modules to improve reusability between the modules. Domain modules are representing sub-domains of the C3PO domain. Table 2 presents the module, their namespace, and the color used in the Figures of Section 4. We published the competency questions for Plant, CropManagment and Plot module on the documentation, and the SPARQL queries.[11,12]

4.1. Ontology modules and knowledge graph

4.1.1. Support modules

Time module

The Time module extends the Time Ontology to be able to represent `c3potime:RelativePropertInterval` composed of `time:RelativeInstant`, as shown in Figure 3, while the respective representations of time instants are not placed in a specific year. This is important for representing cultivation dates for any crop that might occur in a different year. An example of a `c3potime:RelativePropertInterval` could be the date interval between two `time:RelativeInstant`: the 16th of September and the following 2nd of February. This allows to create patterns reusable every year. Moreover, it allows sharing of information that could be reused at any time without reference of the year. `time:RelativeInstant` has the data property `c3potime:inRDate` with a type `c3potime:rdate` formating "_Y_M_F_W_D" where "_" represent a number, "Y" a year, "M" a month, "F" a fortnight, "W" a week, and "D" a day. The previous example should be represented as "0Y9M16D" for the 16th of September and "1Y2M2D" for the 2nd of February.

Vocabulary module

The Vocabulary module corresponds to several closed lists of qualitative values represented as SKOS thesaurus. The SKOS thesaurus has several top concepts dedicated to a specific list: units, climate characteristics, culture modes, etc. Any `skos:Concept` instances are linked to instances from other C3PO modules using specific object properties. For example, the instance of the class `c3poparam:Parameter` is linked to any narrower concepts of `c3povoc:Unit` using the property `c3poparam:hasParameterUnit`. Those `skos:Concept` instances representing unit are partially aligned with QUDT (Hodgson et al., 2014) instances, i.e., a knowledge graph representing the various standard quantity kind and unit.

Parameter module

The Parameter module corresponds to a representation of numeric parameters such as weight and volume. The module is composed of a class `c3poparam:Parameter` with a measured, minimun, and maximum value. The class is specialized for each measurement type. Each parameter class have a constraint to specify its unit, defined in the Vocabulary module.

For example, the class `c3poparam:Yield` will store all the yield measurement with the associated unit `c3povoc:YieldUnit` and is defined as follows:

```
c3poparam:Yield rdf:type owl:Class ;
  rdfs:subClassOf c3poparam:Parameter ,
    [ rdf:type owl:Restriction ;
      owl:onProperty c3poparam:hasParameterUnit ;
      owl:allValuesFrom [owl:intersection
(skos:Concept
          [ rdf:type owl:Restriction ;
            owl:onProperty skos:broader ;
            owl:hasValue c3povoc:YieldUnit]) ;
    rdf:type owl:Class]
  ] .
```

4.1.2. Domain modules

Plant module

The Plant module represents cultivated plant taxonomy from a farmers' viewpoints. Plants are described by the class `c3poplant:CultivatedPlant` with characteristics such as crop seasons, watering needs, or nutritional requirements. Plants are hierarchically organized under a taxonomy representing as a SKOS thesaurus. This taxonomy has different levels as follows: plant family, cultivated plant, varietal type, and cultivar, as shown in Figure 4. Varietal type is an intermediate level between cultivated plant and cultivar that represents some physical characteristics of crops that farmers refer to while defining the market outlets. This level does not belong to a botanical scientific taxonomy. Again using the onion example of the previous section, the botanical and usage families of the onion are, respectively, Amaryllidaceae

[11] https://www.elzeard.co/ontologies/c3po/

[12] https://gitlab.com/serre-des-savoirs/c3po

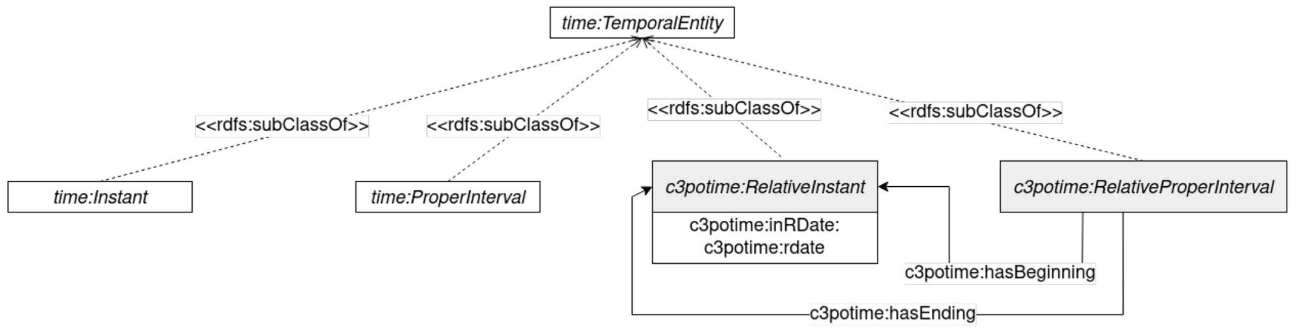


FIGURE 3
Time module representation.

and bulb vegetables. We defined several varietal types such as yellow onions and red onions. Fiamma is an example of a red onion cultivar. The module also represents crop succession and association information. Figure 4 presents the main classes of the Plant module instantiated with the onion example. A Plant knowledge graph was built under the Plant module with the help of agronomists and is linked with FCU and TAXREF-LD. In future, the module will be extended to improve the representation of trees, which will be useful with regard to fruit crops.

Plot module

The Plot module represents the spatial organization of plots on the farms. It contains the representation of a `c3poplot:ProductionCell`, which is an occupation of legally registered land with an address and an id. In addition, the farmer-driven spatial organization is represented with the possibility of creating `c3poplot:CultivablePlot` and `c3poplot:CultivableBed` within the plots, which may be useful for diversified vegetable crop farmers growing multiple crops on the same plot or the same row. Irrigation systems and landscape elements such as meadows present on farms are also represented. Figure 5 presents the main Plot module classes.

In future, the module will be extended to improve the representation of irrigation systems and landscape elements by taking into account all the specific features of the lands, especially in areas of ecological interest that are required to the farm to get different certifications.

CropManagement module

The CropManagement module represents technical itineraries and farming processes for task planning and recording. As noted previously, we studied the three ValueFlows representation layers: plan specification, plan and plan execution. Plan specification is built by instantiating the `c3pocm:CropItinerary` class, representing generic technical itineraries created by farmers and agronomic experts. A `c3pocm:CropItinerary` consist of a set of tasks. One generic task is represented as a `c3pocm:TechnicalOperation`. These generic technical itineraries are linked to a `c3poplant:CultivatedPlant` from the plant module and have parameters such as season and soil type. `c3pocm:TechnicalOperation` are farming processes such as planting or harvesting, which are described with `c3potime:RelativeProperInterval` to give a time range where the task could be applied. Plan is built by instantiating the `c3pocm:ProductionProcess` class for each crop. `c3pocm:ProductionProcess` is composed of a set of `c3pocm:OperationalTask` planned in the farmer's calendar. Farmers often rely on series principles, i.e., they grow the same type of crops with (more or less) the same set of tasks, but on a different plot and at a different time to achieve a continuous flow of crop production. As shown in Figure 6, we integrate the `c3pocm:Series` classes to represent this principle. Plan execution is built by instantiating the `c3pocm:Activity` class. An `c3pocm:Activity` represents one task carried out. `c3pocm:OperationalTask` and `c3pocm:Activity` are tasks that happened on crops and land, so we add a property named `c3pocm:concernsPosition` to link a task and the instance of the `c3poplot:LandUse` class from Plot module. As an example to present the difference between `c3pocm:OperationalTask` and `c3pocm:Activity`, a possible instanciation of `c3pocm:OperationalTask` could be a harvest happened between 4 July 2022 and 6 July 2022, with a certain estimated yield. During the execution, an instantiation of `c3pocm:Activity` is made for each day (4 July 2022, 5 July 2022, and 6 July 2022), with the real yield obtained per day. The three types of task `c3pocm:TechnicalOperation`, `c3pocm:OperationalTask` and `c3pocm:Activity` are linked to instance of `c3pocm:FarmingPractice`. `c3pocm:FarmingPractice` is specialized in many sub-classes, representing various farming tasks such as harvesting or planting. Each sub-class has its own parameters. The three layers help farmers to analyze their production and decide what should be changed the following year to improve their productivity. Figure 6 presents the main classes of the CropManagement module.

The CropManagement knowledge graph built under the C3PO CropManagement module has been partially populated with the help of agronomists, especially regarding the integration of instances of `c3pocm:CropItinerary`.

In future, the module will be extended to link the harvested crops with the Sales module.

Admin module

The Admin module represents agents and organizations and their administrative information as users of the applications. It

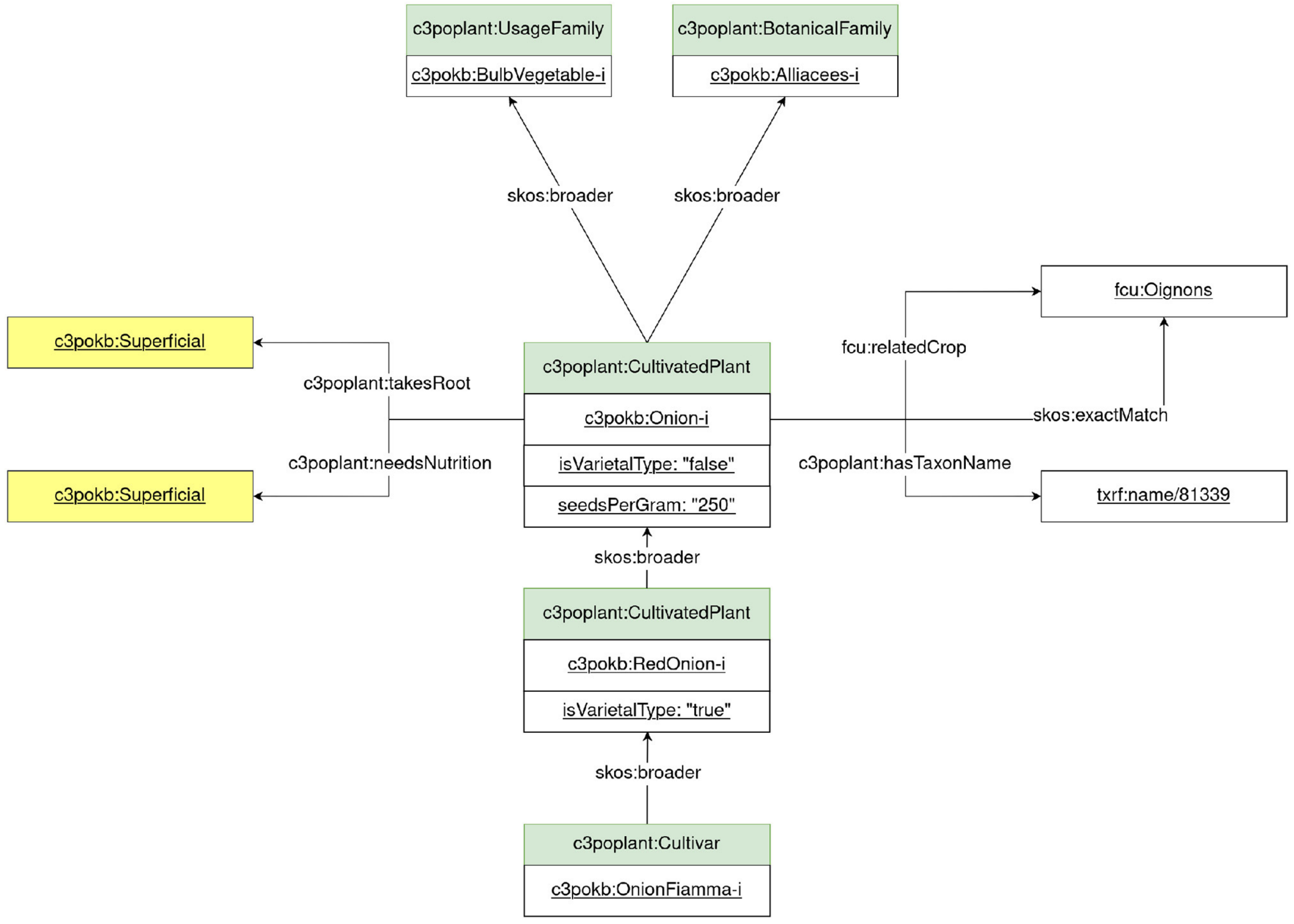


FIGURE 4
Plant module representation.

relies on existing standard ontology modules or vocabularies, such as FOAF (Graves et al., 2007) for the representation of people and organizations and Event Ontology (Raimond and Abdallah, 2007) for events. We created subclasses from these resources to integrate `c3poadmin:Farm`, `c3poadmin:Producer` and `c3poadmin:Cooperative` representations.

Supply module

The Supply module represents agricultural input and equipment that farmers could use. The module allows users to represent input as `ephy:Intrant` and their usages as `ephy:Usage`. A `ephy:Usage` is the combination of a product, plant, or family, a targeted pest and the application method of the product. This combination defines properties such as the maximum dosage allowed or the number of possible applications. Figure 7 presents the main classes of the CropManagement module. The Supply module currently extends the E-PHY ontology to integrate Basagri data. The E-PHY ontology is an ontology produced to represent the French E-Phy catalog of plant protection products. Basagri is a private dataset containing information regarding regulatory data on agricultural inputs in France. The dataset is proposed as a set of files in CSV format. We implemented a pipeline to transform the CSV format into an RDF knowledge graph under the E-PHY ontology as it fulfilled our requirements regarding agricultural inputs. We extracted different information such as the dosage authorized for a product regarding a plant or the number of days required before the farmers return to the plot or harvest, from the Basagri files. We, then, built URIs for input using their marketing authorization (AMM, a code delivered France for authorized chemical products) as in the E-PHY proposition. We created vocabularies using SKOS thesaurus representations for closed lists such as the type of product function (insecticide, herbicide, etc.). We also compared the list of crops of Basagri and C3PO to find similarities based on labels and connect `ephy:Usage` to instance of `c3poplant:CultivatedPlant` or `c3poplant:CultivatedFamily` from the plant module. We used the Levenstein distance (Levenshtein et al., 1966) to deal with slight differences such as singular/plural names. In future, the module will be extended to improve the representation of farming equipment. Moreover, the alignment of Basagri and C3PO crops will also be improved to enhance the number of similarities between the two databases.

Sale module

The Sale module organizes the stocks and the product delivery. The module is still under construction and will be updated and combined with the DataFoodConsortium ontology which

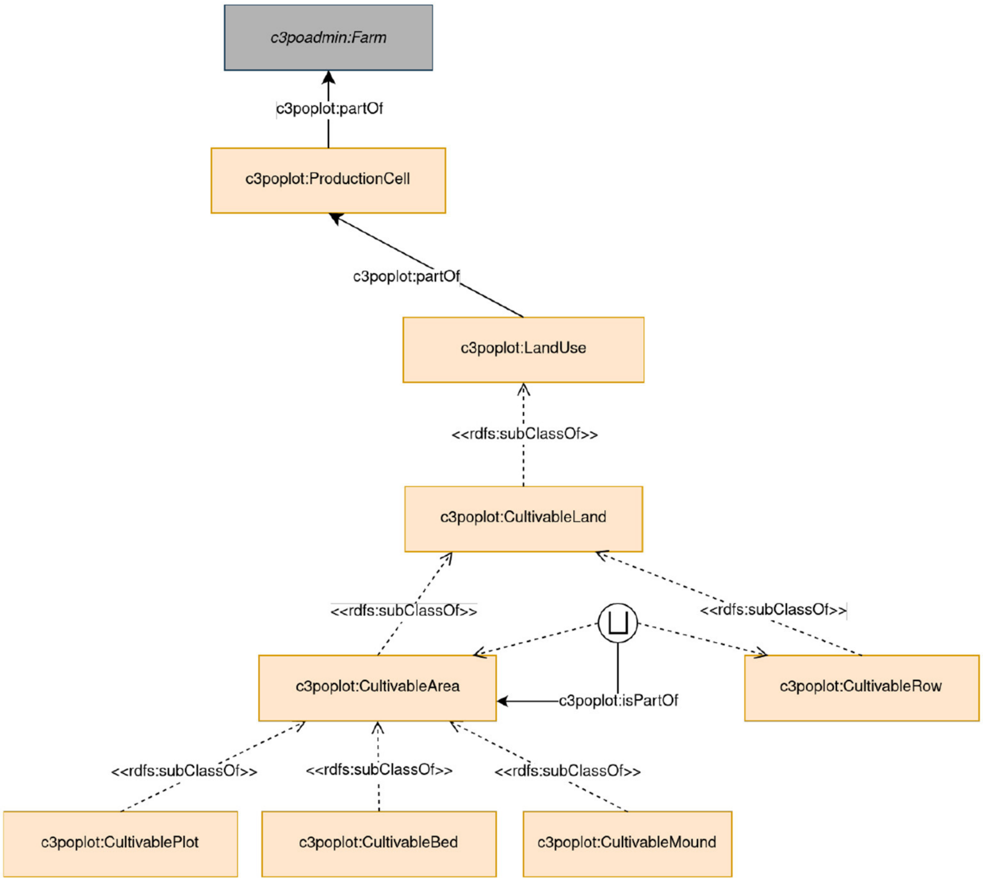


FIGURE 5
Plot module representation.

represents the supply chain and delivery process in the food distribution system.[13]

4.2. Statistics and availability

The C3PO knowledge graph currently consists of 4,647 axioms, 236 classes, 211 object properties, 71 data properties, and 270 individuals mostly contained in the Vocabulary module. Moreover, the KG is currently composed of 8,402,495 triples, of which 3,025,790 are explicit and 5,376,790 are implicit. All of the ontologies are available under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). Figure 8 presents different components of C3PO's knowledge graph (module, inter-module relations, and data source).

A sub-part of the knowledge graph containing the Plant module and the CropManagement module is available on GitLab under an Attribution-ShareAlike 4.0 International license (CC BY-SA 4.0). The other parts are also not public because they concern users' data or are licensed, e.g., Basagri.

According to the MIRO guidelines (Matentzoglu et al., 2018), hereafter in Table 3, we present the current version (1.0) of C3PO. Only the MIRO "basic" guidelines are reported here, but we have incorporated as much metadata as possible in the C3PO OWL source file, according to the MOD specifications (Dutta et al., 2015).

C3PO and its modules are uploaded on AgroPortal. We published each module as a view of the ontology in AgroPortal project. We edited some metadata on the global level: C3PO. We improve the FAIR score of the ontology by following the AgroPortal guidelines during the Metadata AgroHackathon in August 2022.[14]

5. The ontology and knowledge graph in use

We are developing the ontology and knowledge graphs in the context of multiple application development for knowledge sharing and crop planning. These applications are built by Elzeard and developed in the framework of the MESCLUN DURAB, PACON (Morel et al., 2023), and D2KAB (Aubin et al., 2019) research projects.[15] These applications help to assess the ontology

[13] https://www.datafoodconsortium.org/en/

[14] https://agrohackathon2022.workshop.inrae.fr/

[15] https://www.picleg.fr/Projets/Les-projets-en-cours/MESCLUN-DURAB

10.3389/frai.2023.1187090

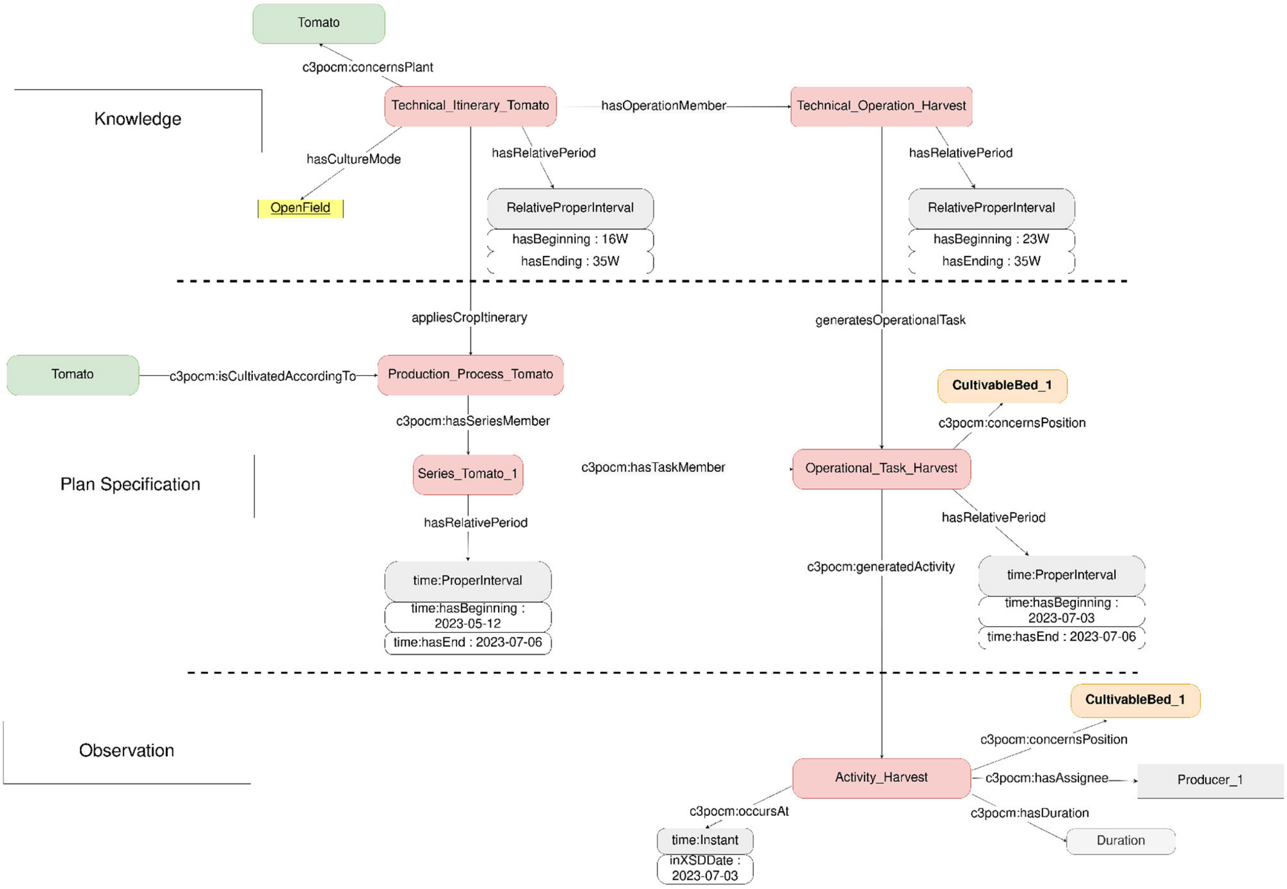


FIGURE 6
CropManagement module representation.

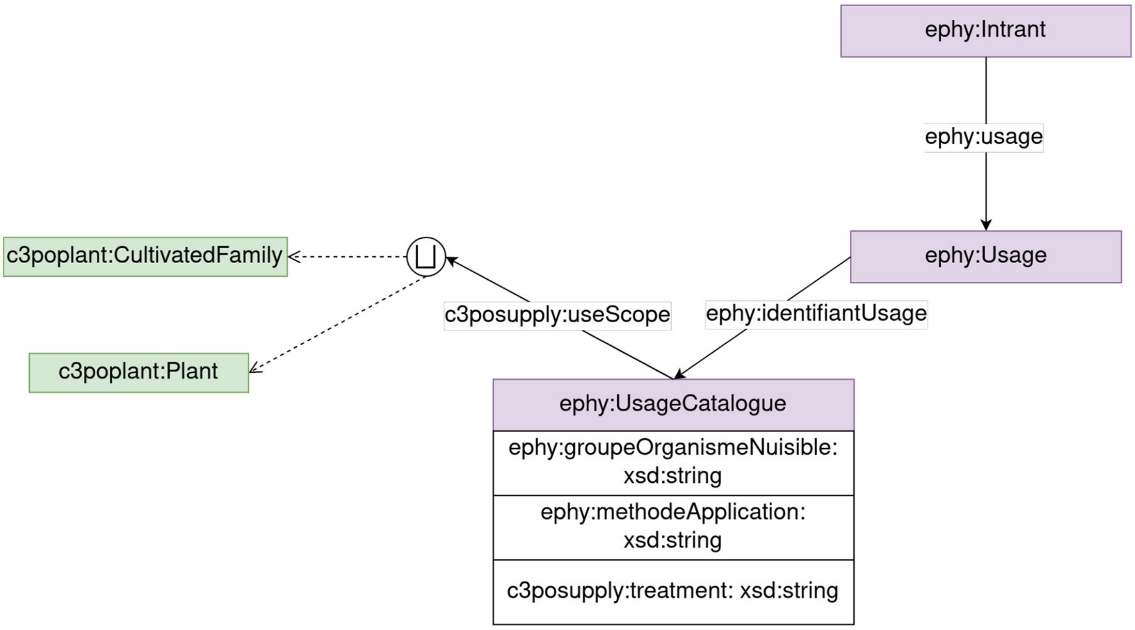

ephy:Intrant
ephy:usage
c3poplant:CultivatedFamily
⨆
ephy:Usage
c3posupply:useScope
ephy:identifiantUsage
c3poplant:Plant
ephy:UsageCatalogue
ephy:groupeOrganismeNuisible: xsd:string
ephy:methodeApplication: xsd:string
c3posupply:treatment: xsd:string

FIGURE 7
Supply module representation.

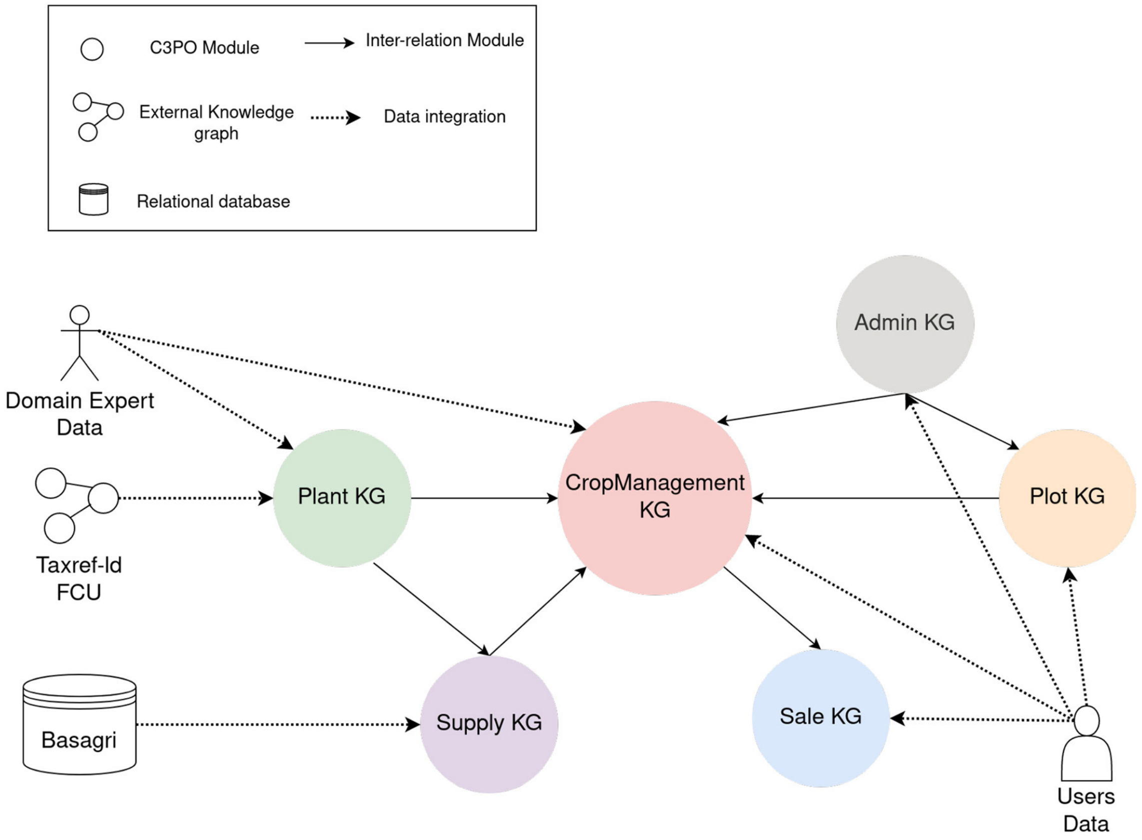


FIGURE 8
Components of C3PO's knowledge graph.

TABLE 3 C3PO information following basics MIRO guidelines.

Basics MIRO guidelines	C3PO information
Admin	http://www.elzeard.co/ontologies/c3po/admin
A.1 Ontology name	Crop planning and production process ontology
A.2 Ontology owner	• Baptiste Darnala (Elzeard) • Florence Amardeilh (Elzeard)
A.3 Ontology license	Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)
A.4 Ontology URL	http://www.elzeard.co/ontologies/c3po
A.5 Ontology repository	https://gitlab.com/serre-des-savoirs/c3po https://agroportal.lirmm.fr/ontologies/C3PO

consistency regarding the domain and the quality of the data integrated in the knowledge graph.

5.1. "*Serre des Savoirs*"

The so called "*Serre des Savoirs*" web portal is under construction to access open data in the C3PO KG related to cultivated plants and technical itineraries. The knowledge graph content is described with the Plant and CropManagement modules. For the CropManagement module, the application will query only the CropItinerary and TechnicalOperation instances. The web portal will directly query the knowledge graph and make it accessible for non-Semantic Web experts such as farmers and agronomists. The development of the application has impacted the development of the Plant and CropManagement module and the needs in terms of information required in the application.

A screenshot of the descriptive page of the tomato is presented in Figure 9. The information displayed provides a general description of this vegetable including the species scientific name, the cultivating families, and varietal types described in Section 3.2. Moreover, information on the cultivation context is provided, such as the irrigation and nutrition needs of the plant. Several competency questions of plant module are used to build this page presented as follows: [16]

1. What is the plant's botanical species?
2. What is the plant's botanical family?
3. What is the scientific name of the botanical taxon (species or family)?
4. What is the plant's usage family?
5. What are the varietal types of a plant?
6. How much does the seed of a plant weigh (seeds per gram)?
7. How much water does a plant need?

[16] https://www.elzeard.co/ontologies/c3po/EN/EN_Plant.html

8. How deep does a plant's root system go?
9. How much nutrients does a plant need?
10. How long does it take for a plant to return to the plot?

To illustrate the querying of the C3PO KG, the SPARQL queries corresponding to CQ2 and CQ10 are presented as follows:

```
PREFIX c3poplant: <http://www.elzeard.co/ontologies/c3po/plant#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

select ?CultivatedPlant ?BotanicalFamily where {
  ?CultivatedPlant a c3poplant:CultivatedPlant .
  ?BotanicalFamily a c3poplant:BotanicalFamily .
  ?CultivatedPlant skos:broader+
?BotanicalFamily . }
```

```
PREFIX c3poplant: <http://www.elzeard.co/ontologies/c3po/plant#>
PREFIX c3poparam: <http://www.elzeard.co/ontologies/c3po/parameter#>

select ?CultivatedPlant ?PlotReturnTime
?parameterValue ?minValue ?maxValue
?ParameterUnit where {
  ?CultivatedPlant c3poplant:hasPlotReturnTime
?PlotReturnTime .
  Optional {?PlotReturnTime
c3poparam:parameterValue ?parameterValue .}
  Optional {?PlotReturnTime c3poparam:minValue
?minValue .}
  Optional {?PlotReturnTime c3poparam:maxValue
?maxValue .}
  ?PlotReturnTime c3poparam:hasParameterUnit
?ParameterUnit . }
```

5.2. Decision support applications

In addition, the C3PO KG was used to build two applications for farmers: *Elzeard* and *La Pépinière*. *Elzeard* is a web application where farmers describe their farms with their locations and plot organization. The farmers can, then, build the plan for their crops and related tasks, organize their farm workers' schedules, and choose the inputs to use. C3PO is used as the data model for the web application. The knowledge graph built with plant, technical itineraries, and input knowledge is queried to help farmers access decision-support information. C3PO is used in *Elzeard*. Figure 10 presents a screenshot of a technical itinerary in *Elzeard*. The list of competency questions involved in building this webpage come from the CropManamgement module:[17]

1. How long does it take for a plant to emerge in this technical itinerary?
2. How long does a plant grow in this technical itinerary?
3. How long does it take to harvest a plant in this technical itinerary?
4. What is the shelf life of a plant in this technical itinerary?
5. What is the estimated overall workload for this technical itinerary?
6. What are the tasks involved in this technical itinerary?
7. When is the best time to plant in this technical itinerary?
8. Over what period will I be able to spread out the harvests for this technical itinerary?
9. Which task must be carried out before or after another task in this technical itinerary?
10. What is the forecast yield for this technical itinerary?
11. Which varieties are recommended for this technical itinerary?
12. What is the expected yield for this technical itinerary?

To illustrate the querying of the C3PO KG, the SPARQL query corresponding to CQ7 is presented as follows:

```
PREFIX c3pocm: <http://www.elzeard.co/ontologies/c3po/cropManagement#>
PREFIX c3potime: <http://www.elzeard.co/ontologies/c3po/time#>
select ?CropItinerary ?FarmingPractice
?beginningDate ?endingDate
where {
  ?CropItinerary a c3pocm:CropItinerary .
  ?CropItinerary c3pocm:hasOperationMember
?Operation .
  ?Operation c3pocm:implements ?FarmingPractice .
  ?FarmingPractice a c3pocm:PlantingProcess.

  Optional {
    ?Operation c3pocm:hasRelativePeriod ?Period .
    ?Period c3potime:hasBeginning ?Beginning .
    ?Beginning c3potime:inRDate ?beginningDate .

    ?Period c3potime:hasEnding ?Ending .
    ?Ending c3potime:inRDate ?endingDate .
  }
}
```

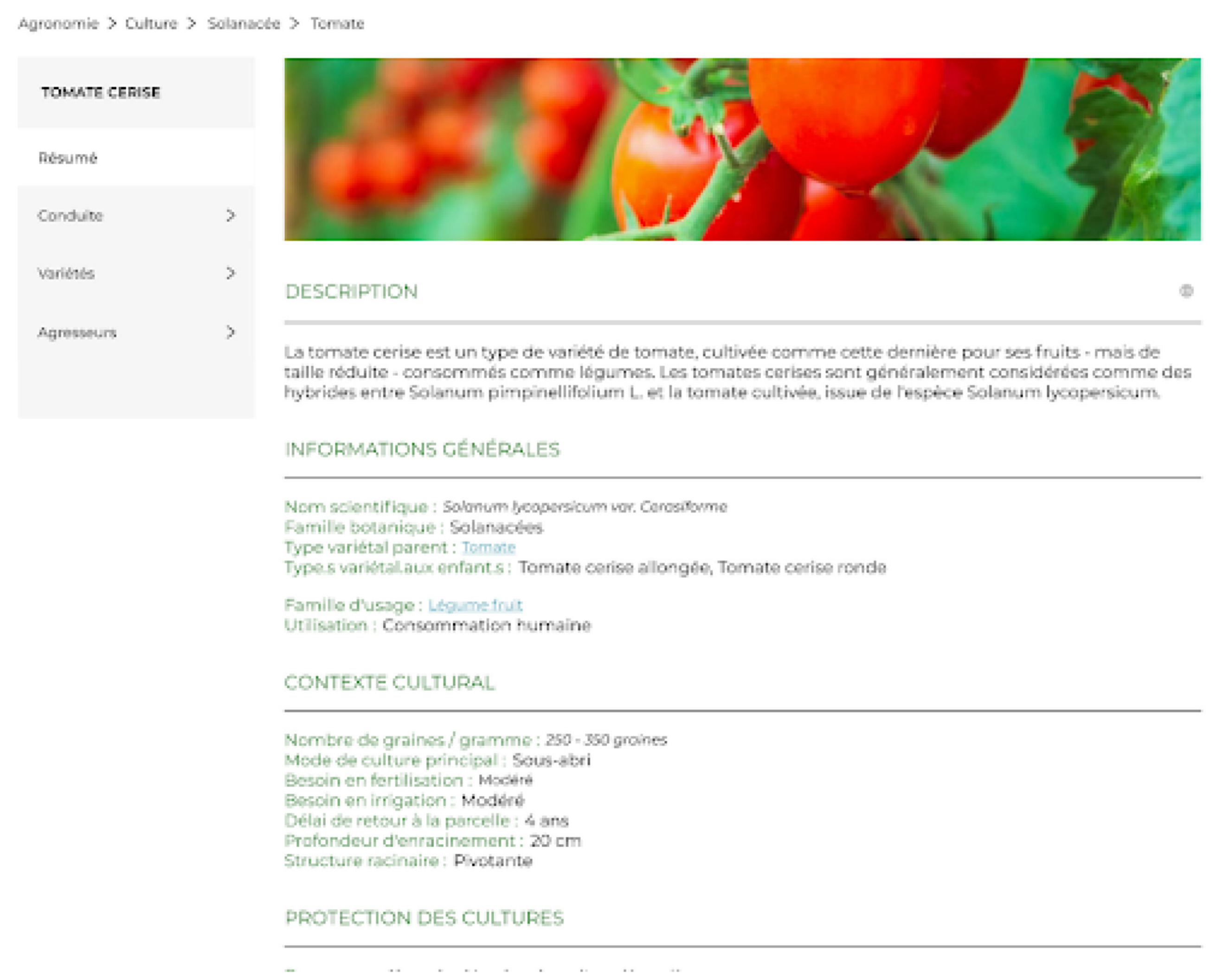
Figure 11 presents a screenshot of a planning made by a farmer, both made in the application *Elzeard*. Figure 10 is composed of several pieces of information such as the cropping period and the cultivating tasks. Figure 11 gives an overview of all the crops of a farmer in a period and the commercial needs in terms of harvested crop. The list of competency questions involved in building this webpage come from the CropManagement module:[18]

1. How many series have I planned for this crop?
2. What is the period of each series that I have planned for this crop?
3. Which variety is associated with this series?
4. What is the surface area associated with this series?
5. What are the planting distances between my seedlings or plants for this series?

La Pépinière is an application under development to help beginner farmers design their farms and future productions. The application has an educative feature, whereas *Elzeard* is production-oriented purpose. *La Pépinière* uses the Plant, CropManagement, and Plot modules as data models and has access to the same knowledge present in *La Serre des Savoirs* to help farmers. Figure 12 presents a screenshot of the *La Pépinière* application. The competency questions used are the same as presented in Figure 11.

[17] https://www.elzeard.co/ontologies/c3po/EN/EN_CM.html

[18] https://www.elzeard.co/ontologies/c3po/EN/EN_CM.html



Agronomie > Culture > Solanacée > Tomate

TOMATE CERISE

Résumé

Conduite >

Variétés >

Agresseurs >

DESCRIPTION

La tomate cerise est un type de variété de tomate, cultivée comme cette dernière pour ses fruits - mais de taille réduite - consommés comme légumes. Les tomates cerises sont généralement considérées comme des hybrides entre Solanum pimpinellifolium L. et la tomate cultivée, issue de l'espèce Solanum lycopersicum.

INFORMATIONS GÉNÉRALES

Nom scientifique : *Solanum lycopersicum var. Cerasiforme*
Famille botanique : Solanacées
Type variétal parent : Tomate
Type.s variétal.aux enfant.s : Tomate cerise allongée, Tomate cerise ronde

Famille d'usage : Légume fruit
Utilisation : Consommation humaine

CONTEXTE CULTURAL

Nombre de graines / gramme : *250 - 350 graines*
Mode de culture principal : Sous-abri
Besoin en fertilisation : Modéré
Besoin en irrigation : Modéré
Délai de retour à la parcelle : 4 ans
Profondeur d'enracinement : 20 cm
Structure racinaire : Pivotante

PROTECTION DES CULTURES

FIGURE 9
Descriptive page of the tomato in *La Serre des Savoirs*.

6. Discussion and difficulties

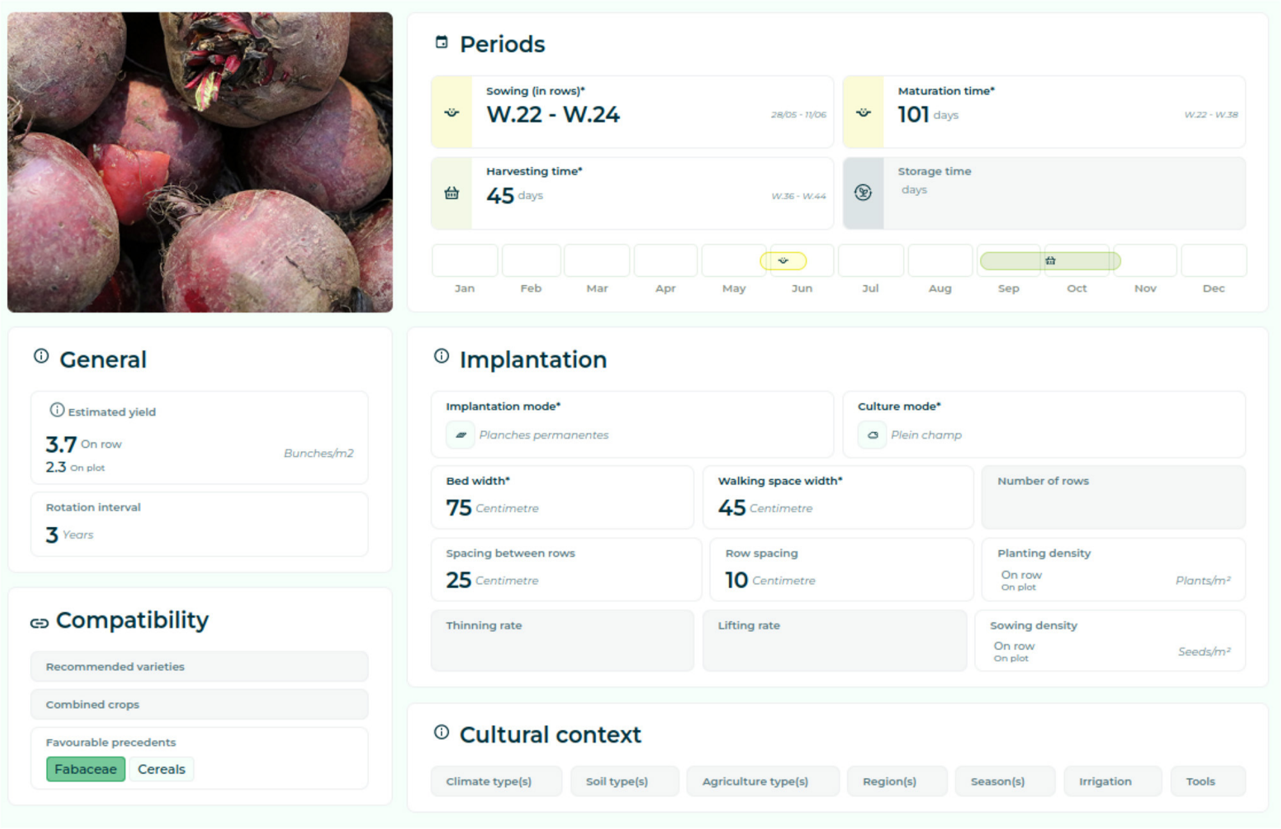
We encountered several difficulties during C3PO development process. Concerning C3PO development, the diversified vegetable agricultural domains and the multiple viewpoints were complex to represent in a single ontology. The division into modules eases the development by allowing to work in multiple subdomains independently. However, naming problems of classes shared between multiple modules could arise and should be checked when updating the ontology. In addition, this complexity required the involvement of multiple domain experts such as farmers, agronomists, taxonomists, and retailers to have a better vision of the domain and the needs the ontology should meet.

As illustrated, C3PO, even if not yet perfect, was developed primarily to serve multiple applications, and thus, a usable ontology had to be produced quickly. The Agile development methodology, which eases the release of several iterative ontology versions already usable by application developers, was suitable. However, new domain discoveries could impact previous development choices and lead to refinement of the ontology impacting the knowledge graph structure. Moreover, our methodology helps us to keep track all steps of development process. The files (Chowlk schema, text description,...) produced during the specification and conceptualization steps are used to document C3PO.

Regarding cooperation with the development team and domain experts, we had to use tools that are easily understandable by non-Semantic Web users. The graphic notation proposed by Chowlk helped to produce schemes that would be understandable by different actors and directly convertible into OWL format, but this generated a more complex ontology engineering workflow, otherwise we would have simply used Protégé in group.

In the documentation writing process, we included definitions of classes and properties in shared documents to improve collaboration between DEs and OEs. It was important to provide understandable detailed definitions to enhance the model reusability.

During the C3PO conceptualization process, various difficulties were encountered according to the concerned module. For the



Periods

Sowing (in rows)*: W.22 - W.24 — 28/05 - 11/06

Maturation time*: 101 days — W.22 - W.38

Harvesting time*: 45 days — W.36 - W.44

Storage time: days

General

Estimated yield: 3.7 On row, 2.3 On plot — Bunches/m2

Rotation interval: 3 Years

Compatibility

Recommended varieties

Combined crops

Favourable precedents: Fabaceae, Cereals

Implantation

Implantation mode*: Planches permanentes

Culture mode*: Plein champ

Bed width*: 75 Centimetre

Walking space width*: 45 Centimetre

Number of rows

Spacing between rows: 25 Centimetre

Row spacing: 10 Centimetre

Planting density: On row, On plot — Plants/m²

Thinning rate

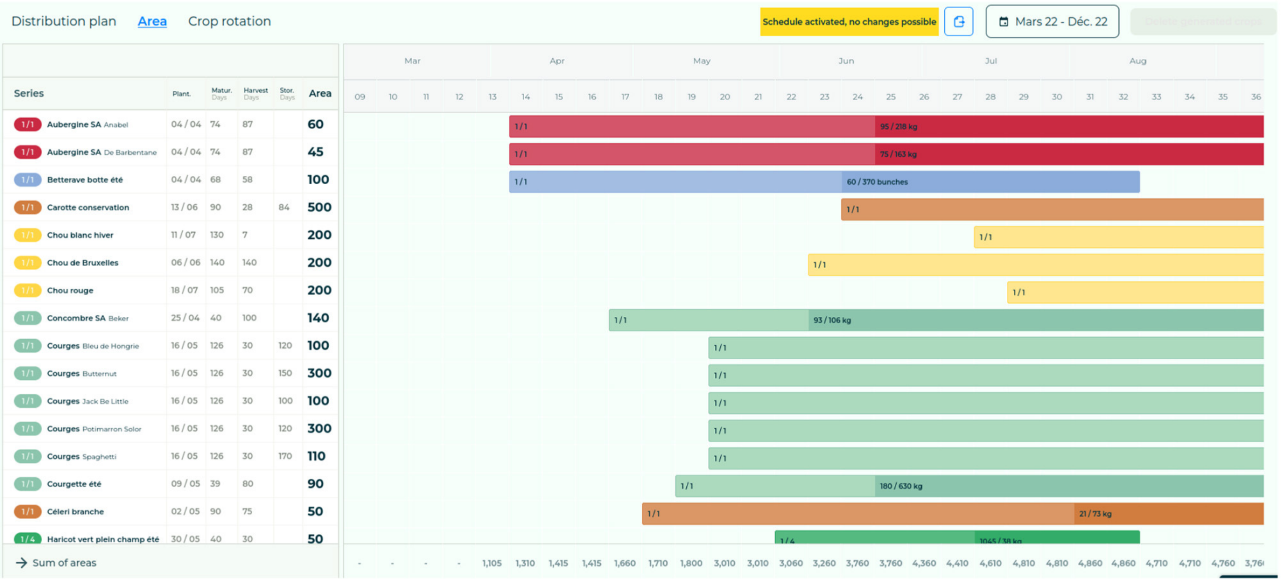
Lifting rate

Sowing density: On row, On plot — Seeds/m²

Cultural context

Climate type(s) | Soil type(s) | Agriculture type(s) | Region(s) | Season(s) | Irrigation | Tools

FIGURE 10
Technical itinerary of the beet in *Elzeard*.



Distribution plan | Area | Crop rotation

Schedule activated, no changes possible

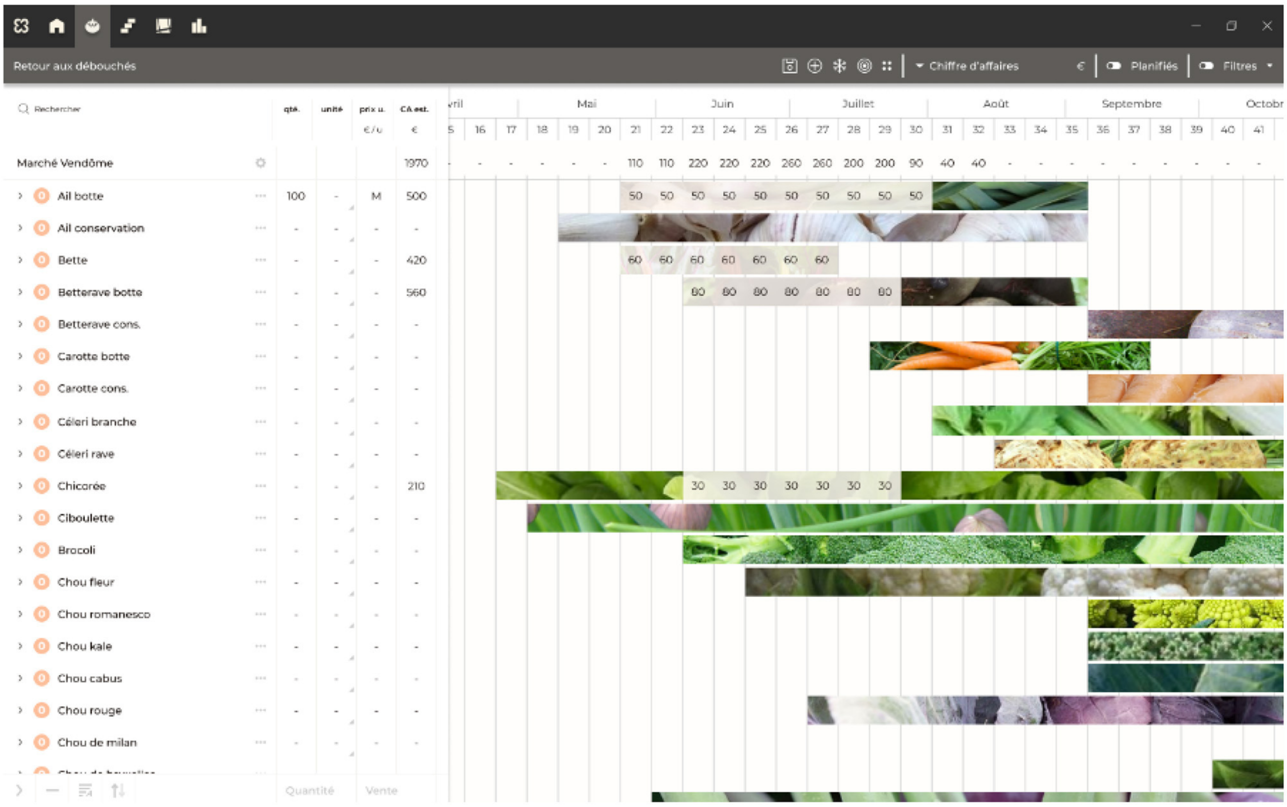
Mars 22 - Déc. 22

Series	Plant.	Matur. Days	Harvest Days	Stor. Days	Area
1/1 Aubergine SA Anabel	04 / 04	74	87		60
1/1 Aubergine SA De Barbentane	04 / 04	74	87		45
1/1 Betterave botte été	04 / 04	68	58		100
1/1 Carotte conservation	13 / 06	90	28	84	500
1/1 Chou blanc hiver	11 / 07	130	7		200
1/1 Chou de Bruxelles	06 / 06	140	140		200
1/1 Chou rouge	18 / 07	105	70		200
1/1 Concombre SA Beker	25 / 04	40	100		140
1/1 Courges Bleu de Hongrie	16 / 05	126	30	120	100
1/1 Courges Butternut	16 / 05	126	30	150	300
1/1 Courges Jack Be Little	16 / 05	126	30	100	100
1/1 Courges Potimarron Solor	16 / 05	126	30	120	300
1/1 Courges Spaghetti	16 / 05	126	30	170	110
1/1 Courgette été	09 / 05	39	80		90
1/1 Céleri branche	02 / 05	90	75		50
1/4 Haricot vert plein champ été	30 / 05	40	30		50

	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
→ Sum of areas	-	-	-	-	1,105	1,310	1,415	1,415	1,660	1,710	1,800	3,010	3,010	3,060	3,260	3,760	3,760	4,360	4,410	4,610	4,810	4,810	4,860	4,860	4,710	4,710	4,760	3,76

FIGURE 11
Crop planning of a farmer in *Elzeard*.

Plant module, our aim was to create a module that represents a plant taxonomy with multiple viewpoints (botanical and agricultural), to have agricultural information on plants and ease the link with different heterogeneous knowledge graphs. This led to the typing of instances as `skos:Concept` and `owl:Class`. SKOS offers the possibility of creating the plant organization and



	qté.	unité	prix u. €/u	CA est. €
Marché Vendôme				1970
Ail botte	100	-	M	500
Ail conservation	-	-	-	-
Bette	-	-	-	420
Betterave botte	-	-	-	560
Betterave cons.	-	-	-	-
Carotte botte	-	-	-	-
Carotte cons.	-	-	-	-
Céleri branche	-	-	-	-
Céleri rave	-	-	-	-
Chicorée	-	-	-	210
Ciboulette	-	-	-	-
Brocoli	-	-	-	-
Chou fleur	-	-	-	-
Chou romanesco	-	-	-	-
Chou kale	-	-	-	-
Chou cabus	-	-	-	-
Chou rouge	-	-	-	-
Chou de milan	-	-	-	-

	5	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
Marché Vendôme	-	-	-	-	-	-	110	110	220	220	220	260	260	200	200	90	40	40	-	-	-	-	-	-	-	-	-
Ail botte							50	50	50	50	50	50	50	50	50	50											
Bette							60	60	60	60	60	60	60														
Betterave botte									80	80	80	80	80	80	80												
Chicorée									30	30	30	30	30	30	30												

FIGURE 12
Crop planning of a farmer in *La Pépinière*.

links with other `skos:Concept`, while OWL helps to create classes and properties. For the Supply module, the transformation of an existing ontology (E-PHY) led to checking if the ontology was usable and the changes required to meet our needs. The most important changes were made when we opted to change the domain or the range of a property. In that case, we chose to create a new property. The main drawback of this method is the E-PHY ontology, and its properties are provided in French while C3PO is built in English, so the URIs of the module are not in the same languages. Linguistic uniformization could be applied in future.

During the C3PO KG building process, multiple and heterogeneous sources of data required us to build multiple data integration pipelines. For the Plant module, we had to test different data integration scenarios with domain experts. As previously noted, we ended by using tabular files. This method eases the integration of domain expert data but could lead to problems. As multiple spreadsheets are used, problems of misspelled URIs can occur and lead to missing connections in the graph. We overcame this problem by creating a list of SPARQL queries to check the consistency of the graph, find errors, and fix them before importing the data in the graph. We also had difficulty connecting our instances with TAXREF-LD as taxons are represented with `owl:Class`, and scientific names are represented by instances of `skos:Concept`. We decide to link our C3PO instances with `skos:Concept` using specific C3PO's properties. Notably, a link between a class and an instance is possible only with the property `rdf:type`. Regarding the data imported from the applications, problems were encountered due to wrong data integration or failed knowledge graph updates. For instance, properties were duplicated instead of being renamed after an ontology update. Here, we used SPARQL queries to address these issues. We recommend to update the knowledge graph by exporting the data, applying the change, and importing in a new knowledge graph. Thus, we keep track of changes to be able to roll back. In this way, the knowledge graph is not updated directly.

About the reasoning aspect, we define some constraints on C3PO's classes to check the quality of users' input data (`c3poparam:Parameter`). However, we do not check the expert data extracted from reference sources (TAXREF-LD, FCU, etc.). Unfortunately, mistakes may happen on this source that will cause inconsistencies. Thus, we should apply reasoning and other checking processes on those part of the graph in future.

The knowledge graph is not fully opened. A subpart regarding information about plants and technical itineraries is accessible in GitHub and through a SPARQL endpoint. User information saved in the Plot, Admin and CropManagement modules remain private. Sharable information will be accessible through the "La Serre des Savoirs" Web Portal. This portal will be enriched with data already aggregated, and users will have the possibility to enrich "La Serre des Savoirs" directly.

Regarding the interoperability and reusability challenges, we applied different workflows. First, we linked a subpart of our plant instances with other knowledge graphs (FCU and TAXREF-LD). However, we recreate the concepts, we aggregate a subpart of the knowledge such as labels, and we keep the alignment. We process as well to control the management of the terminology and to enrich it with multiple sources. In addition, we are using core domains ontologies such as FOAF, Prov-O, or Time ontology. Regarding future development to reuse or align with domain ontologies, we prefer to align our concepts instead of import external concepts. This choice is made regarding the context of industrial development, as we cannot ensure that external ontologies will be sustainable. Align instead of import offer the possibility to keep the control on C3PO. In addition to that, major concepts such as technical itineraries do not exist in other ontologies, which reduce the possibility of reusing this part of the graph. Finally, domain ontologies reused in C3PO are stored in AgroPortal repository and were found through this repository. Thus, ontologies not declared in the repository were not studied during our conceptualization step. However, we may miss some interesting ontologies such as PestOn (Medici et al., 2022), which means that we should update regularly our state-of-the-art research.

7. Conclusion and future work

Diversified vegetable farming is complex, and many parameters have to be taken into account for decision support. We built the Crop Planning and Production Ontology (C3PO) and its Knowledge Graph to help farmers in their choices. The ontology is divided into several modules to represent a specific part of the domain. The knowledge graph is created from heterogeneous data sources (other knowledge graphs, relational databases, or user/expert data). The C3PO KG is the backbone of three web applications and aims to give farmers access to information to support their planning and monitoring decisions. The open part of the knowledge graph brings novel aspect as no representation of technical itineraries exists for vegetables farmers. This knowledge has not been formalized yet and serve as a basis for reusability of common technical itineraries shared in different sources. Future studies will be continued to the referencing and sharing of technical itineraries to create a collaborative knowledge base through our web portal “La Serre des Savoirs” currently in development.

These applications–not all yet in production–already pre-validate C3PO as an “application ontology”, but future reuses will also validate C3PO as a “domain ontology”. The methodology presented in this study is based on LOT and SAMOD methodologies, and we highlighted how we implemented each process in an application development operation. Various domain expert partners were included in our approach to assess and identify the main concepts and properties: scientists (from the D2KAB, MESCLUN DURAB research project) and agricultural professionals (crop farmers, networks of agricultural advisors, and teachers). C3PO is available on GitLab as an open source project that can be reused and contributed to and published in AgroPortal to facilitate its discovery and reuse. In future studies, we will extend and improve the ontology to include equipment, farm components (e.g., irrigation structures or meadows currently present in the ontology but need refinement), and pests and diseases. We will also improve the ability of the ontology to make inferences on the data based on agricultural knowledge. We will improve the interoperability of C3PO and create alignment with other semantic resources. Finally, we will extend the scope of the ontology and knowledge graph in order to be able to model other types of crop production, such as arboriculture or agroforestry.

Data availability statement

TAXREF-LD: The version 15.2 of TAXREF-LD graph used for this study can be found in the AgroPortal repository https://agroportal.lirmm.fr/ontologies/TAXREF-LD. The github repository is https://github.com/frmichel/taxref-ld. The SPARQL EndPoint is https://taxref.mnhn.fr/sparql. FCU: The version 3.3 of the FCU thesaurus used for this study can be found in the AgroPortal repository: https://agroportal.lirmm.fr/ontologies/CROPUSAGE. The gitlab repository is https://gitlab.irstea.fr/copain/frenchcropusage. The SPARQL EndPoint is http://ontology.inrae.fr/frenchcropusage/sparql. C3PO KB: The version 1.0 of the C3PO KB can be found in the gitlab repository https://gitlab.com/serre-des-savoirs/c3po-kb. The associated ontology can be found on the AgroPortal repository https://agroportal.lirmm.fr/ontologies/C3PO/?p=summary. The SPARQL EndPoint is https://graph.elzeard.co/sparql.

Author contributions

BD supervised the building of C3PO and its KG and used within applications and wrote the manuscript. BD and FA built the ontology, the knowledge graph, and implemented the process with domain experts. BD, FA, and CR wrote the definitions of the ontology. CJ and CR helped in the modeling of some modules of the ontology and provided general knowledge engineering expertise. All authors contributed to the manuscript, read, and approved the final version.

Funding

The authors would like to acknowledge support from the National Office for Biodiversity with MesclunDurab grant and I-NOV grant. This study was also partially achieved with support from the *Data to Knowledge in Agronomy and Biodiversity* (D2KAB - www.d2kab.org) project that received funding from the French National Research Agency (ANR-18-CE23-0017) and the project Partages de Connaissances (PACON) of the transverse program MetaBio funded by INRAE. This work also received fund from Agence Nationale de la Recherche under the France 2030 program, reference ANR-16-CONV-0004.

Acknowledgments

The authors would like to thank Dr. Kevin Morel (INRAE), Juliette Raphel (Elzeard), and Guillaume Turlier (Elzeard) for help as domain experts and all contributors from MesclunDurab and D2KAB projects for their constructive feedback.

Conflict of interest

BD and FA were employed by Elzeard, Bordeaux.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Amdouni, E., Bouazzouni, S., and Jonquet, C. (2022). O'FAIRe makes you an offer: metadata-based automatic FAIRness assessment for ontologies and semantic resources. *Int. J. Metadata, Semantics Ontologies*. 16, 16–46. doi: 10.1504/IJMSO.2022.131133

Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R., Matteis, L., et al. (2012). "Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes," in *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (Barcelona), 220–225. doi: 10.5220/0004138302200225

Arora, N. K. (2019). Impact of climate change on agriculture production and its sustainable solutions. *Environm. Sustainab.* 2, 95–96. doi: 10.1007/s42398-019-00078-w

Aubin, S., Adam-Blondon, A.-F., Alaux, M., Ba, M., Bernard, S., Bisquert, P., et al. (2019). "D2KAB project taking off: Data to Knowledge in Agronomy and Biodiversity," in *RDA P14 2019 - 14th Research Data Alliance Plenary Conference* (Helsinki). doi: 10.5281/zenodo.352030

Bouazzouni, S., and Jonquet, C. (2021). "L'ontologie e-phy, une base de connaissances pour le catalogue des produits phytopharmaceutiques autorisés en agriculture en France," in *Journées Francophones d'Ingénierie des Connaissances (IC) Plate-Forme Intelligence Artificielle (PFIA'21)*, 105–112.

Chávez-Feria, S., García-Castro, R., and Poveda-Villalón, M. (2022). "Chowlk: from uml-based ontology conceptualizations to owl," in *The Semantic Web: 19th International Conference, ESWC 2022, Hersonissos, Crete, Greece, May 29-June 2, 2022, Proceedings*. Cham: Springer, 338–352.

Darnala, B., Amardeilh, F., Roussey, C., Todorov, K., and Jonquet, C. (2022) "Ontological representation of cultivated plants: linking botanical and agricultural usages," in *MK 2022 - 1st Workshop on Modular Knowledge @ ESWC 2022* (Hersonissos), 165–173.

Devare, M., Aubert, C., Laporte, M.-A., Valette, L., Arnaud, E., and Buttigieg, P. (2016). "Data-driven agricultural research for development: a need for data harmonization via semantics," in *Proceedings of the Joint International Conference on Biological Ontology and BioCreative, Corvallis, Oregon, United States*.

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., et al. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci. Food* 2, 23. doi: 10.1038/s41538-018-0032-6

Drury, B., Fernandes, R., Moura, M.-F., and de Andrade Lopes, A. (2019). A survey of semantic web technology for agriculture. *Infor. Proc. Agricult.* 6, 487–501. doi: 10.1016/j.inpa.2019.02.001

Dumont, B. (2021). Why working conditions are a key issue of sustainability in agriculture? A comparison between agroecological, organic and conventional vegetable systems. *J. Rural Stud.* 56, 53–64. doi: 10.1016/j.jrurstud.2017.07.007

Dutta, B., Nandini, D., and Shahi, G. K. (2015). "Mod: metadata for ontology description and publication," in *International Conference on Dublin Core and Metadata Applications* (São Paulo), 1–9. Available online at: https://dcpapers.dublincore.org/pubs/article/view/3758

Garijo, D. (2017). "Widoco: a wizard for documenting ontologies," in *The Semantic Web-ISWC 2017: 16th International Semantic Web Conference, Vienna, Austria*. Cham: Springer, 94–102.

Glimm, B., Horrocks, I., Motik, B., Stoilos, G., and Wang, Z. (2014). HermiT: An OWL 2 reasoner. *J. Autom. Reason.* 53, 245–269. doi: 10.1007/s10817-014-9305-1

Graves, M., Constabaris, A., and Brickley, D. (2007). Foaf: connecting people on the semantic web. *Cat. Classif. Q.* 43, 191–202. doi: 10.1300/J104v43n03_10

Hobbs, J. R., and Pan, F. (2006). Time ontology in owl. *W3C Working Draft.* 27, 3–36.

Hodgson, R., Keller, P. J., Hodges, J., and Spivak, J. (2014). *Qudt-Quantities, Units, Dimensions and Data Types Ontologies*. Available online at: http://qudt.org (accessed September 13, 2023).

Isbell, F., Adler, P., Eisenhauer, N., Fornara, D., Kimmel, K., Kremen, C., et al. (2017). Benefits of increasing plant diversity in sustainable agroecosystems. *J. Ecol.* 105, 871–879. doi: 10.1111/1365-2745.12789

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E. A., McCouch, S., Pujar, A., et al. (2005). Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics* 6, 388–397. doi: 10.1002/cfg.496

Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Yeumo, E. D., Emonet, V., et al. (2018). Agroportal: a vocabulary and ontology repository for agronomy. *Comput. Electron. Agric.* 144, 126–143. doi: 10.1016/j.compag.2017.10.012

Larmande, P., and Todorov, K. (2021). "Agrold: A knowledge graph for the plant sciences," in *The Semantic Web-ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event*. Cham: Springer, 496–510.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., et al. (2013). "PROV-O: The PROV Ontology," in *W3C Recommendation*. Cambridge, MA: World Wide Web Consortium.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady* 10, 707–10.

Matentzoglu, N., Malone, J., Mungall, C., and Stevens, R. (2018). MIRO: guidelines for minimum information for the reporting of an ontology. *J. Biomed. Semant.* 9, 6. doi: 10.1186/s13326-017-0172-7

Medici, M., Dooley, D., and Canavari, M. (2022). Peston: An ontology to make pesticides information easily accessible and interoperable. *Sustainability* 14, 6673. doi: 10.3390/su14116673

Mendelsohn, R. (2009). The impact of climate change on agriculture in developing countries. *Nat. Resour. Res.* 1, 5–19. doi: 10.1080/19390450802495882

Michel, F., Gargominy, O., Tercerie, S., and Faron-Zucker, C. (2017). "A model to represent nomenclatural and taxonomic information as linked data. application to the french taxonomic register, TAXREF," in *Proceedings of ISWC 2017 Workshop on Semantics for Biodiversity (S4Biodiv 2017), Oct 2017, Vienna, Austria*, 1–12.

Miles, A., and Bechhofer, S. (2009). "Skos simple knowledge organization system reference," in *W3C Recommendation. World Wide Web Consortium, United States.*

Morel, K., Cerf, M., Jeuffroy, M.-H., Roussey, C., Amardeilh, F., and Appert, P. (2023). PArtage, interopérabilité et mobilisation des COnnaissances par le Numérique pour la (re)conception de fermes biologiques (PACON)," in *French Métabio workshop "changement d'échelle de l'AB"* (Saint Malo). Available online at: https://hal.inrae.fr/hal-04077432

Morel, K., and Léger, F. (2015). "Strategies to manage crop planning complexity in very diversified direct selling farming systems: the example of organic market gardeners," in *Proceedings of the 5th International Symposium for Farming Systems Design*, 93–94.

Musen, M. A. (2015). The protégé project: a look back and a look forward. *AI Matters* 1, 4–12. doi: 10.1145/2757001.2757003

Palma, R., Roussaki, I., Döhmen, T., Atkinson, R., Brahma, S., Lange, C., et al. (2022). "Agricultural information model," in *Information and Communication Technologies for Agriculture-Theme III: Decision*, eds D. D. Bochtis, C. G. Sørensen, S. Fountas, V. Moysiadis, and P. M. Pardalos. Cham: Springer International Publishing, 3–36.

Paut, R., Sabatier, R., and Tchamitchian, M. (2019). Reducing risk through crop diversification: An application of portfolio theory to diversified horticultural systems. *Agricult. Syst*. 168, 123–130. doi: 10.1016/j.agsy.2018.11.002

Peroni, S. (2016). *SAMOD: An Agile Methodology for the Development of Ontologies.* Peterborough: Bytes Publisher, 1579911.

Poveda-Villalón, M., Fernández-Izquierdo, A., Fernández-López, M., and Garcí-a-Castro, R. (2022). LOT: An industrial oriented ontology engineering framework. *Eng. Appl. Artif. Intell.* 111, 104755. doi: 10.1016/j.engappai.2022.104755

Poveda-Villalón, M., Gómez-Pérez, A., and Suárez-Figueroa, M. C. (2014). OOPS! (OntOlogy pitfall scanner!): An on-line tool for ontology evaluation. *Int. J. Semant. Web Inf. Syst*. 10, 7–34. doi: 10.4018/ijswis.2014040102

Raimond, Y., and Abdallah, S. (2007). *The Event Ontology*. Available online at: https://motools.sourceforge.net/event/event.html (accessed September 13, 2023).

Roussey, C. (2018). "Frenchcropusage: Thésaurus sur les cultures françaises," in *le thésaurus décrivant les cultures françaises par leur utilisation au format skos*. Available online at: https://fairsharing.org/ (accessed September 13, 2023).

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., and Katz, Y. (2007). Pellet: A practical OWL-DL reasoner. *J. Web Semant*. 5, 51–53. doi: 10.1016/j.websem.2007.03.004

Soulignac, V., Pinet, F., Lambert, E., Guichard, L., Trouche, L., and Aubin, S. (2019). Geco, the french web-based application for knowledge management in agroecology. *Comp. Electron. Agricult*. 162, 1050–1056. doi: 10.1016/j.compag.2017.10.028

Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., et al. (2015). "Vocbench: a web application for collaborative development of multilingual thesauri," in *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31-June 4, 2015*. Cham: Springer, 38–53.

Suárez-Figueroa, M. C., Gómez-Pérez, A., Fernáandez-López, M. (2012). "The NeOn methodology for ontology engineering," in *Ontology Engineering in a Networked World*, eds M. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, A. Gangemi (Berlin; Heidelberg: Springer). doi: 10.1007/978-3-642-24794-1_2

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scient. Data* 3, 1–9 doi: 10.1038/sdata.2016.18

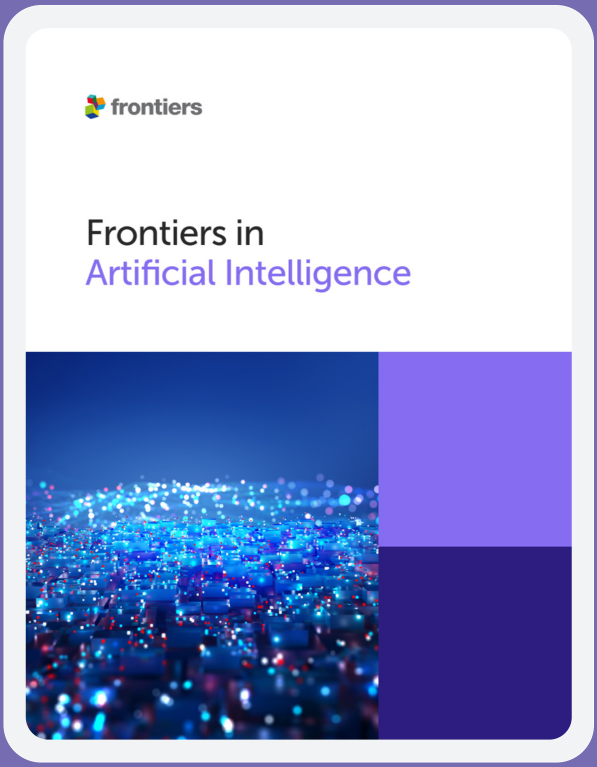
Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Artificial Intelligence

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us
+41 (0)21 510 17 00
frontiersin.org/about/contact