Decision support tool for the agri-food sector using data annotated by ontology and Bayesian network: a proof of concept applied to milk microfiltration.

C Baudrit, Patrice Buche, Nadine Leconte, Christopher Fernandez, Maëllis Belna, Geneviève Gésan-Guiziou

International Journal of Agricultural and Environmental Information Systems
Volume 13 • Issue 1

Decision Support Tool for the Agri-Food Sector Using Data Annotated by Ontology and Bayesian Network:
A Proof of Concept Applied to Milk Microfiltration

Cédric Baudrit, I2M, University of Bordeaux, INRAE, Bordeaux, France
Patrice Buche, IATE, University of Montpellier, INRAE, Institut Agro, Montpellier, France*
Nadine Leconte, STLO, INRAE, Institut Agro, Rennes, France
Christophe Fernandez, I2M, University of Bordeaux, INRAE, Bordeaux, France
Maëllis Belna, Boccard, Research and Development, F-35360 Montauban-de-Bretagne, France
Geneviève Gésan-Guiziou, STLO, INRAE, Institut Agro, Rennes, France

ABSTRACT

The scientific literature is a valuable source of information for developing predictive models to design decision support systems. However, scientific data are heterogeneously structured expressed using different vocabularies. This study developed a generic workflow that combines ontology, databases, and computer calculation tools based on the theory of belief functions and Bayesian networks. The ontology paradigm is used to help integrate data from heterogeneous sources. Bayesian network is estimated using the integrated data taking into account their reliability. The proposed method is unique in the sense that it proposes an annotation and reasoning tool dedicated to systematic analysis of the literature, which takes into account expert knowledge of the domain at several levels: ontology definition, reliability criteria, and dependence relations between variables in the BN. The workflow is assessed successfully by applying it to a complex food engineering process: skimmed milk microfiltration. It represents an original contribution to the state of the art in this application domain.

KEYWORDS

Bayesian Network, Data Integration, INRAE, Knowledge Base, Knowledge Integration, Milk Microfiltration, Ontology, Reliability, Uncertainty

1. INTRODUCTION

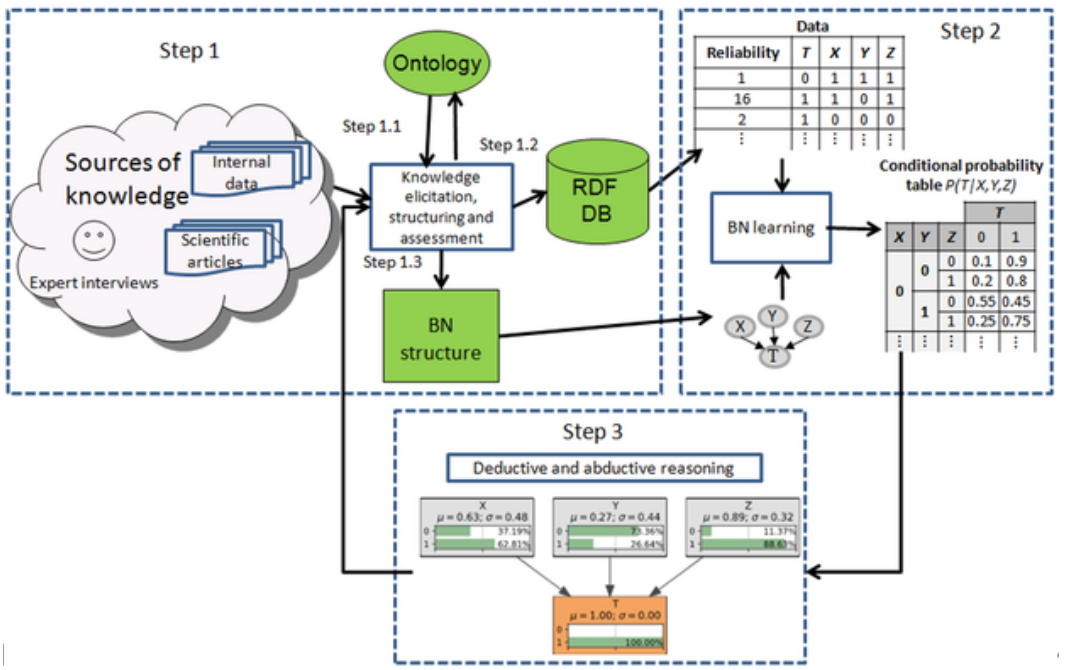
For decision tasks such as optimising food processes, an initial step is to predict variables of interest from process parameters. The scientific literature, including experimental data and knowledge expressed by domain experts, is a valuable source of information to reach this goal. However, the ever-increasing amount of scientific data is heterogeneously structured, found mainly in text format and expressed using different vocabularies. Addressing this difficulty requires innovative tools that

DOI: 10.4018/IJAEIS.309136 *Corresponding Author

can integrate and treat new information. In this context, using Semantic Web methods based upon ontologies seem relevant to structure experimental information (Lousteau-Cazalet et al. 2016; Yeumo et al. 2017; Aubin et al. 2019). As experiments use different methods and technologies, another difficulty is considering source (document) reliability when using the data in calculations. The theory of belief functions provides suitable solutions to address this issue (Destercke et al. 2013). Providing relevant conclusions and recommendations requires developing adequate modelling tools that can integrate, as much as possible, available knowledge which is heterogeneous in nature and quality. Such modelling tools must be able to manage heterogeneous sources of knowledge (experimental data and expert opinion), multiple manipulated scales and different forms of uncertainty (Perrot et al. 2016; Barnabe et al. 2018). With this goal in mind, Bayesian networks (BNs) (Jensen and Nielsen, 2007; Pearl, 1988) provide a practical mathematical structure that can describe complex systems which contain uncertainty. BNs are based on a coupling between graph and probability theory in which the graph provides an intuitively appealing interface with which model designers can represent strongly interacting sets of variables. Uncertainty in the system is considered by quantifying the dependence between variables in the form of conditional probabilities. The use of BNs has been investigated recently in agri-food domains (Baudrit et al. 2015; Drury et al. 2017; Chapman et al. 2018).

This article discusses a numerical workflow to treat data and knowledge that combines ontologies, databases and computer calculation tools based on the theory of belief functions and BNs. The workflow developed is based on a pluridisciplinary collective study involving experts in the domains of food processing and artificial intelligence, and comprises three sequential steps (see Fig. 1). The first step consists of elicitation, structuring and assessment of knowledge related to a food process of interest. More precisely, experimental data published in scientific articles are annotated using an ontology, and their reliability is assessed by experts in food processing. Data from scientific articles are annotated in a simple tabular format file that is semi-automatically generated using the ontology (see step 1.1 in Fig. 1). Then, in step 1.2, the file is uploaded and annotated data are stored in a Resource Description Framework (RDF) database. The complete annotation data set used in this

Figure 1. Workflow process developed in this study. RDF: Resource Description Framework, DB: database, BN: Bayesian network.



paper is available from (Buche et al. 2021) and the database can be queried in open access using a SPARQL Protocol and RDF Query Language (SPARQL) end-point. Relationships between variables from expert opinion are structured by a BN through its associated graph. The second step consists of extracting annotated data and associated reliability scores from the database using a dedicated querying system guided by the ontology to learn the BN parameters (i.e. conditional probability tables). The third step consists of reasoning via inference with the model developed in order to predict process parameters, which is an initial step in optimising the food process and thus the design of a decision support system. This is an iterative workflow which can be enriched with new data and knowledge without damaging the structure of the entire workflow.

The feasibility and utility of the solution focus on the process of skimmed milk microfiltration operating with a membrane of 0.1 μm pore diameter. Milk microfiltration is applied to fractionate milk proteins (casein micelles and serum proteins) in order to produce innovative ingredients. This process remains difficult to understand as a whole, especially because existing models (Astudillo-Castro et al. 2020) assess only a specific range of operating conditions and do not take into account the three used types of membrane technologies (Gésan-Guiziou, 2010): (1) ceramic membranes with uniform transmembrane pressure (UTP), (2) ceramic membranes with graded hydraulic resistance (*eg* GP® membranes) and (3) polymeric spiral wound membranes. The innovations in this article include (1) combining ontologies and BN to structure heterogeneous sources of data/knowledge, given the reliability of data sources, to provide relevant conclusions and recommendations based on deductive and quantitative reasoning; (2) the development of a new domain ontology representing skimmed milk microfiltration and a set of metadata to assess data source reliability; and (3) assessment of the workflow developed (Fig. 1) based on an actual complex application in the domain of skimmed milk microfiltration to answer questions of domain experts.

Section 2 discusses studies that combine the ontology paradigm with probabilistic graphical models and highlights the utility of modelling for microfiltration plants. Section 3 focuses on the system for annotating heterogeneous data sources guided by ontology and the reliability assessment model used to implement step 1. Section 4 includes preliminary ideas on BNs and the process, used in step 2, to learn parameters taking into account the reliability of extracted data. To finish, Section 5 discusses the results obtained using the workflow developed in this study for an optimisation problem in milk microfiltration where a new domain ontology has been designed to annotate the dataset according to the reliability of sources.

2. RELATED STUDIES AND THE UTILITY FOR MICROFILTRATION PLANTS

Ontology is a formal representation of knowledge that structures the domain of knowledge via hierarchical specialisation organisation of concepts and the relationships between them (Staab and Studer, 2010). Supported by a description logic language such as the Web Ontology Language (OWL), ontology-based information systems enable automatic logical reasoning. Multiple approaches have been developed to manage probabilistic reasoning in ontology,(Setiawan et al. 2017). Several languages have emerged to tackle BNs in ontologies such as PR-OWL (Carvalho et al. 2017), BayesOWL (Pan et al., 2005) or OntoBayes (Yang and Calmet, 2005) extending OWL. PR-OWL designs ontologies that contain probabilistic information by considering the properties of OWL classes as random variables; it relies on the framework of Multi Entity Bayesian Networks (Laskey, 2008) which combines the first-order logic principles with BNs. BayesOWL and OntoBayes (its successor) propose an approach to directly express OWL ontologies as BNs where classes are translated into nodes and relations are represented as conditional probabilities. Unfortunately, these frameworks remain prototypical, cannot query in SPARQL and lose the richness of the ontology framework. There is a lack of standardised tools capable of dynamically and automatically updating ontology or supporting BN structural adjustment and parameter learning. This leads to a weak scalability of the aforementioned languages (Patnaikuni et al. 2017). Bayesian ontologies obtained with these approaches cannot integrate all information

contained in the ontology because BNs cannot represent relational information. Thus, the two models must be combined. A variety of approaches have been used to translate ontologies into BNs; for example, Devitt et al. (2006) and Fenz (2012) used different algorithms to automatically generate BNs from existing ontologies.. In order to avoid losing the relational aspect of ontology,. Ishak et al. (2011), proposed an algorithm for transforming an ontology into the form of an object-oriented BN. The method developed by Truong et al. (2005) merged ontology and probabilistic relational models (PRMs), into a new model which supports different types of reasoning. Compared to these studies, the approach detailed in the present study doesn't aim to merge BNs into ontology or to transform ontology into BNs but rather to develop an operational workflow which takes advantage of the richness of ontologies and BNs frameworks.

In a more recent study, PRMs learned from data and the semantic information of the ontology (Munch et al. 2017) before building BNs. Munch et al. (2019) developed a method to identify new causal relationships from an ontology and (Munch et al. 2021) use this framework for decision making in the field of food packaging composite design. Compared to our approach which permits to represent n-ary relations of any domain (Buche et al. 2013), their application domains are restricted to process and observation information. Moreover, their approach is not well suited to data annotation based on a systematic review of the literature contrary to ours, which also permits to assess data source reliability and considers it in the probabilistic reasoning.

The aim of the approach presented in this paper is to be able to extract data and knowledge stemming from heterogeneous literature sources of variable reliability, guided by a domain ontology, in order to supply BNs and inversely. The final objective is to address questions provided by experts, which require reason in the face of uncertainty. Consequently, a simple-to-use generic ontological model was selected which has already been used successfully for food (Guillard et al. 2017, Yun et al. 2018) and bio-based products (Lousteau-Cazalet et al. 2016) to manage heterogeneous literature sources. Representing relations between data in this ontological core model as n-ary relations (Buche et al. 2013) guarantees acceptance by end-users who are familiar with entering and manipulating data in spreadsheets. The core ontology model has been implemented in @Web software (AtWeb 2021) which permits to (1) annotate data from heterogeneous sources in a simple tabular format that is semi-automatically generated by @Web using the ontology and (2) import the annotated tables into the RDF database. Relation concepts in the domain ontology, which are specialisations of the core ontological model, define accurately and without ambiguity the information that must be extracted from the data sources. The issue of data source reliability was addressed, and a simple solution was developed to include it in the BN learning phase. The workflow developed evolves in the sense that it can identify knowledge gaps and be enriched with new knowledge.

To our knowledge, the proposed method is unique in the sense that it proposes an annotation and reasoning tool dedicated to systematic analysis of the literature, which takes into account expert knowledge of the domain at several levels: definition of ontology, reliability criteria and dependence relations between variables of interest in the BN.

Milk microfiltration is becoming increasingly attractive in the dairy sector (Gésan-Guiziou et al. 1999; Saboya et al. 2000; Brans et al. 2004). Crossflow microfiltration with a 0.1 μm membrane is widely used to separate the native casein micelles (~150 nm) from serum proteins (~2-15 nm) within skimmed milk. The casein concentrated retentate is used to standardise milk prior to cheese making. The permeate, which contains serum proteins, is further ultra-filtered to provide protein-rich concentrates with high nutritional and functional properties. The increasing interest in these milk protein fractions explains the expansion of milk microfiltration equipment in the dairy sector. Despite this growing interest, the need remains to control the process, improve prediction of microfiltration performances and optimise plant design. Microfiltration includes three membrane technologies (Gésan-Guiziou, 2010): (1) ceramic membranes with uniform transmembrane pressure (UTP) (Sandblöm, 1974), which consists in the circulation of the permeate to obtain a homogeneous transmembrane pressure (TMP) along the membrane; (2) ceramic membranes with a graded resistance (e.g. GP® or ISOFLUX®

membranes) (Skrzypek and Burger, 2010) and (3) polymeric spiral wound membranes. Predicting the performance of these systems accurately is difficult (and sometimes impossible) because modelling of the microfiltration process is rare due to the lack of information about the transport phenomena involved in the microfiltration process and their influence on the performance of the process. When models exist (e.g. Astudillo-Castro (2015, 2020)), they assess only one membrane technology with one milk pre-treatment under a specific range of operating conditions that does not necessarily correspond to the industrial range of operating parameters. The choice of membrane technology and associated operating conditions, as well as the overall design of the filtration equipment implanted industrially, are based mainly on the knowledge of each equipment manufacturer (Belna et al. 2020). However, the databases of these equipment manufacturers do not contain enough data on fraction composition, operating parameters or plant design to accurately predict the performances of microfiltration under a wide range of operating conditions. To our knowledge, they do not use information in the scientific literature. This lack of data integration makes it difficult to predict performances and compare the membrane technologies available on the market, and hinders the optimisation of microfiltration plants. Some studies have focused on developing computational fluid dynamics approaches to predict and improve the performance of microfiltration. For example, Jalivand et al. (2014) simulated crossflow microfiltration of whey suspensions by solving Navier-Stockes equations combined with Darcy's law. However, these approaches have not been able to predict well the microfiltration of skimmed milk under a wide range of membrane technologies and processing conditions.

Consequently, the originality of this study in the field of food engineering is to apply the proposed method based on ontology and BN to predict performances of milk microfiltration under a wide range of operating conditions and membrane technologies.

3. HETEROGENEOUS DATA SOURCE ANNOTATION AND RELIABILITY ASSESSMENT

Using ontologies is one relevant solution to integrate heterogeneous sources of scientific data (Buron et al. (2020)). Although relational models are widely popular and have been used since the 1980s for storing and retrieving data, their dependence on a rigid schema and the explosive growth of available data result in the reduction of their interest and importance. Indeed, the schema of a relational database makes it difficult to add new relationships between the objects or to manage the interoperability between different sources of databases. Implementing complex systems in relational databases requires introducing associative tables (also known as join tables) when many-to-many relationships occur in the model and this is expensive to be calculated. To overcome these limits, RDF graph-oriented models propose a versatile framework capable of flexible extensions to take into account new sources of information. Moreover, contrary to relational databases, multiple RDF databases may be queried simultaneously on the Web thanks to federated SPARQL queries that facilitate interoperable data reuse in an Open Science perspective. Finally, domain ontologies, which may be natively used in RDF databases to annotate data, promote the use of aligned and standardized vocabularies, which reduce ambiguities in human and machine understanding.

@Web tool has been used to implement the first step of Fig. 1. This includes semantic annotation of data from scientific articles guided by an ontology (Buche et al. (2013)), data source reliability assessment (Destercke et al. (2013)) and querying of the annotated data stored in a RDF database available on the Web. @Web is based on a generic ontological and terminological resource (OTR) which is used to annotate and query scientific data . OTR is divided in two parts: a generic part (i.e. core ontology) and a specific part targeted to a given domain of application (i.e. domain ontology). As the core ontology lies at the heart of the workflow of capitalising on scientific data, only the domain ontology must be determined to reuse @Web in a new application domain. By example, Lousteau-Cazalet et al. (2016) applied it to the biomass process domains. The core ontology is structured to model scientific experiments in a given domain in annotated data tables. It is indeed a simple solution, easily

understandable by annotators to structure an experimental result comprising the observed phenomenon and the associated relevant parameters, which are modelled by n-ary relation concepts. Relation and Argument are generic concepts defined in the core ontology to model n-ary relations and arguments related by these relations. Concepts in a given application domain are defined as specialisations in the concepts of the core ontology. The set of terms that describe the application domain in different languages and which is associated with concepts in the conceptual portion used to annotate data is represented in the terminological part of the OTR. More details on the structure of the OTR can be found in (Buche et al. (2013)). @Web is fully compliant with semantic web standard languages:the ontological part of OTR is modelizedin OWLthe terminological part in SKOS, annotated tables in RDF, and the querying module based on SPARQL.

When collecting data from multiple sources, the reliability of both data and scientific articles rapidly becomes an issue. @Web includes a tool to estimate reliability (Step 1 in Fig. 1), which (Destercke et al. (2013)) describes in detail. Given a scientific article *a* retrieved from literature databases on the Internet, @Web tool provides an interval score $\left[\underline{E_a}, \underline{E_a}\right]$ that reflects the *a priori* (i.e., avoiding specific examination) reliability of information provided in the article. The score is computed using *metadata*, and the amplitude of $\left[\underline{E_a}, \underline{E_a}\right]$ provides information about the consistency of the multiple metadata taken into account. Interested readers by the design of the system are invited to read (Destercke et al. (2013). In the @Web querying module, interval scores are exploited to rank according to their reliability annotated data tables associated with articles. They have also been used to assess the reliability of eco-efficient indicators associated with innovative transformation processes in the biorefinery domain (Lousteau-Cazalet et al. (2016)).

4. BAYESIAN NETWORK MODELLING FOR THE DECISION SUPPORT TOOL

A Bayesian network(BN) (Jensen and Nielsen, 2007; Pearl, 1988) is a compacted representation of a joint multivariate probability distribution over a set of random variables. The graph depicts the structure of the BN where the arcs capture properties of probabilistic conditional independence between variables. The nodes of BN correspond to random variables containing probabilistic information. A conditional probabilistic table is assigned to each node (known as the parameters of BN) expressing the probability of its associated variable given its parent nodes in the graph. The joint probability distribution over all node values can be expressed as the product of these local conditional probabilities given as:

$$P\left(X_1,\cdots,X_n\right)=\prod_{i=1}^{n}P\left(X_i \mid Pa\left(X_i\right)\right) \tag{1}$$

where $P\left(X_i \mid Pa\left(X_i\right)\right)$ is the conditional probability function associated with random variable X_i, conditioned on its parents Pa(X_i). The variables in BN may be discrete, continuous or a combination of both. Only discrete networks were considered in the present study.

The structure and parameters of BN may be obtained either by expert elicitation or by machine learning from substantial and/or incomplete data or both (Heckerman, 2008; O'Hagan, 2006, Ji et al. (2015)). In this study, the structure of a graph is provided by expert opinion guided by the ontology. Let θ_{ijk} be the probability that $X_i = x_k$, given that its parents have instantiation x_j :

$$\theta_{ijk}=P\left(X_i=x_k \mid Pa\left(X_i\right)=x_j\right) \quad i=1,\cdots,n \quad j=1,\cdots,c_i \quad k=1,\cdots,r_i \tag{2}$$

where r_i is the number of values that random variable X_i can take, and c_i is the number of distinct configurations of $Pa\left(X_i\right)$. Baudrit et al. (2013) developed a hybrid method which learns to estimate BN parameters from multiple sources of knowledge. Parameters θ_{ij} are initialised by using Dirichlet prior distributions and are successively updated each time new information is made available and can be formulated into a frequentist form. This approach is used in step 2 of the workflow (Fig. 1) and enables weighting the importance of the multiple sources of knowledge, as follows:

$$\theta_{ijk} \mid \left(D_1,\cdots,D_m\right) = \frac{\alpha_{ijk} + \sum_{p=1}^{m} s\left(p\right) N_{ijk}\left(p\right) / N_{ij[k]}\left(p\right)}{\alpha_{ij[k]} + \sum_{p=1}^{m} s\left(p\right)} \tag{3}$$

where [.] denotes a sum, $\alpha_{ij} = (\alpha_{ij1},\cdots,\alpha_{ijr_i})$ are the hyperparameters of the Dirichlet prior distribution, which can be thought of as the size of a virtual database which corresponds to a belief of experts rather than on experiments. Database D_p can correspond to simulated or experimental data, where counts $N_{ij}\left(p\right)$, and $s\left(p\right)$ are the confidence level (reliability rank) of source of knowledge p. In this study, the reliability rank is estimated from the scores established in section 3 and $N_{ij}\left(p\right)$ corresponds to the events in the dataset extracted from @Web.

Using BNs, known as inference, consists of estimating the conditional probability $P(X_Q \mid X_E)$ of a set of variables X_Q given a set of evidence variables X_E. For further details about inference, see Salmeron et al. (2018), who show different kinds of inference algorithms (exact and approximate inference) based on the complexity and size of the BN (Cooper, 1990).

5. WORKFLOW APPLICATION FOR SKIMMED MILK MICROFILTRATION

To demonstrate the workflow developed (Fig. 1) in the context of food processing, the study focused on the hydraulic performance of milk microfiltration with ceramic membranes and did not consider protein transmission or the quality of fractions.

As no ontology of milk microfiltration was found on existing portals or in the literature, experts designed a new domain ontology. By using *Competency Questions* (Uschold & Gruninger 1996, natural-language questions that outline the scope of knowledge represented by an ontology), the milk microfiltration ontology has been designed and adapted to the @Web OTR format and is available on AgroPortal and the INRAE dataverse (MICROFILTRATION ontology). It currently contains 20 relation concepts involving 152 symbolic concepts and 73 quantity concepts associated with 49 units of measurement.

Fig. 2 shows an n-ary relation Relation among dynamics of controlled microfiltration parameters, which models operation of the microfiltration unit and connects an input product – milk that has undergone a series of pre-treatments (e.g. skimming, heating, removal of bacteria) – to operating parameters associated with microfiltration, such as the permeation flux (Jp). This n-ary relation represents a dynamic process. The time parameter associates a timestamp with a given observation during microfiltration.

Fig. 3 presents an extract of the OTR structure composed of its core component and the milk microfiltration. It may be noticed that the Argument concept is specialised in Quantity concept and Symbolic concept in the core ontology. Time (resp. Input product), which belongs to the milk microfiltration domain ontology is a specialisation of Quantity (resp. Symbolic) concept. Time and

Figure 2. An excerpt of the n-ary relation relation among dynamics of controlled microfiltration parameters to model operation of the microfiltration unit (called microfiltration_controlled_parameter_evolution_relation in the OTR)

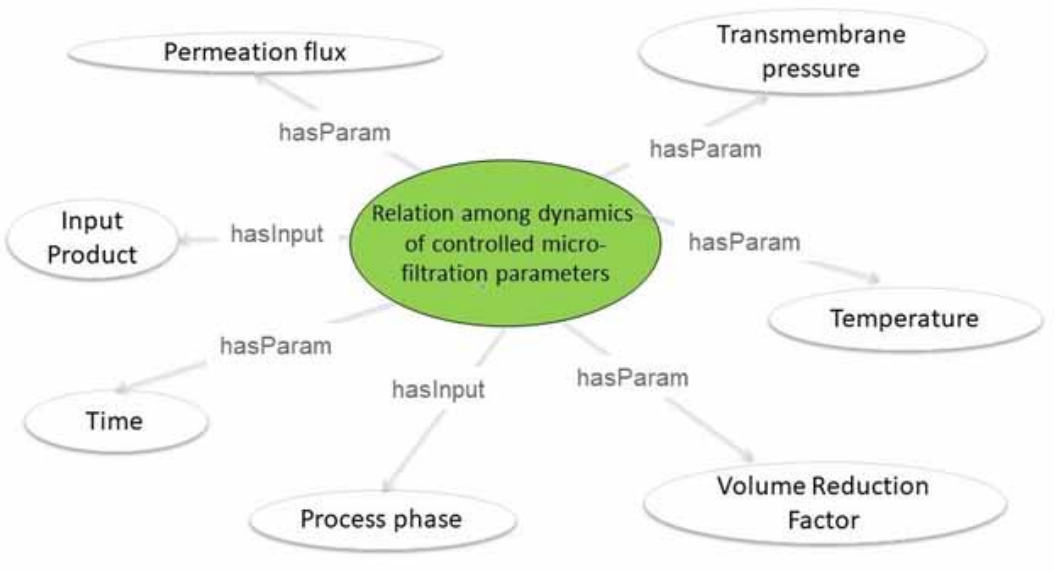
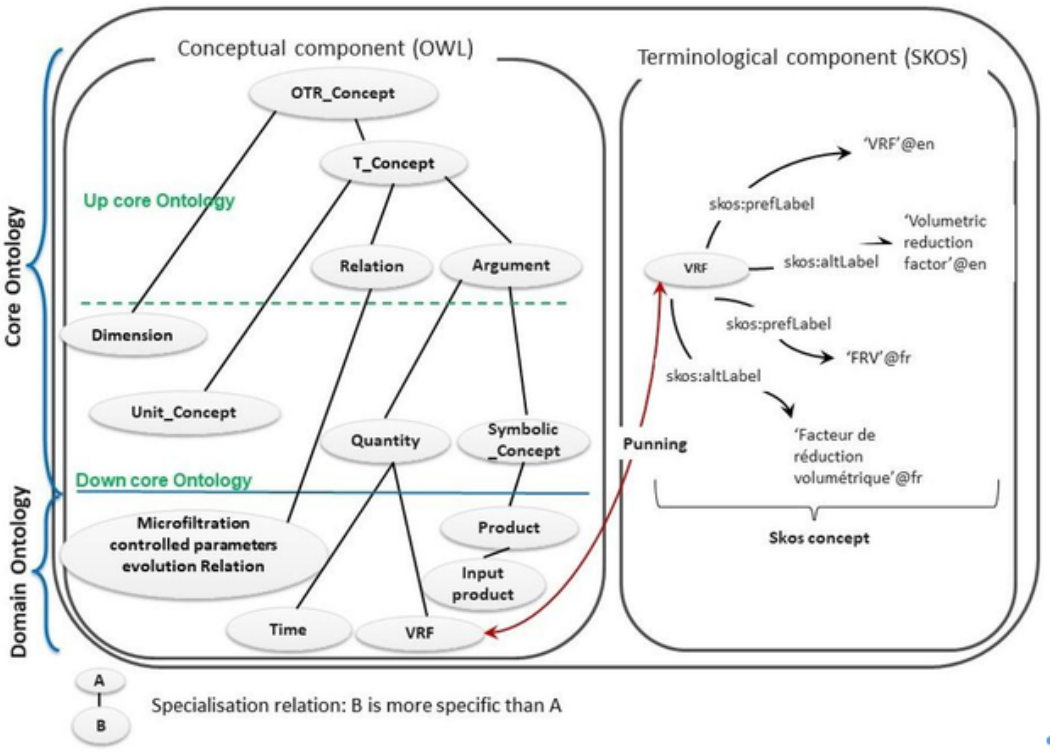


Figure 3. An excerpt of the ontological and terminological resource (OTR) model specialised for milk microfiltration (SKOS: Simple Knowledge Organization System)



Input product are arguments of the microfiltration_controlled_parameter_evolution_relation presented in Fig. 2.

The database created with the milk microfiltration OTRconsists of 52 scientific articles in the domain of milk microfiltration, for a total of 220 documents. Other articles were excluded because they performed microfiltration using innovative methods (e.g. gas injection, rotating membranes) that did not reflect traditional microfiltration. Several of the articles initially identified studied the influence of microfiltration on the cheese making but did not give enough details about operating conditions. Other documents did not provide the required characterization of product and permeate and retentate fractions. The data included in the set of 52 selected articles represent:

- 190 process experiments
- 1572 controlled variables at several sampling times
- 731 product characterisations

Two excerpts of annotated tables which include data from experiments of Jorgensen et al. (2016) and conform to the milk microfiltration ontology are shown below. The structure, in terms of columns, of the annotated tables, has been generated automatically by @Web using a selection of relation concepts previously defined in the OTR. In order to homogenise data structuration extracted from heterogeneous sources, the annotator is invited to find in the article the information requested in each column. Once fulfilled, annotated tables which conform to the OTR may be automatically imported into the RDF database thanks to the import @Web module.

The first table (Table 1) shows one experiment involving five relation concepts in the ontology associated with unit operations commonly used in milk microfiltration (see column Treatment). For example, Experiment 1 is composed of a sequence of five unit operations. The second unit operation

Table 1. Excerpt from the annotated table process description for milk microfiltration (MF), including data from Jorgensen et al. (2016)

Treatment	Input product	Expe no.	Process step no.	Treatment duration	Treatment duration_UNIT	Tempe-rature	Temperature_UNIT	Membrane reference	Manufacturer	Membrane material
Skimming	Product input separation step	1	1			63	°C			
Heat Treatment	Product input separation step	1	2	15	s	73	°C			
Storage	Product input separation step	1	3	0	h	4	°C			
Temperature Holding	Product input separation step	1	4	15	min	45	°C			
MF Separation Micelles/ Serum Proteins	Product input separation step	1	5					not specified	ORELIS	Titania-Zirconia

(column named *Process step number* which equals 2), which is a heat treatment, shows the controlled parameters of treatment duration (15 seconds) and temperature (73°C). The fifth unit operation is the operation of the microfiltration unit, which includes several characteristics of the membrane (e.g. manufacturer, membrane material).

An excerpt of the annotated table *Dynamics of controlled microfiltration parameters* (Table 2), which involves one relation concept of the ontology (Fig. 2), corresponds to the fifth unit operation (Microfiltration Separation Micelles/Serum proteins) of Experiment 1 (Table 1). Values associated with the controlled parameters (e.g. VRF, TMP, Jp) were annotated at different sampling times. Examples of RDF data presented in Table 2 and a SPARQL query to retrieve the subset of columns (Experience_number, ProcessStep_number, Time, VRF) are given in Annex 1.

The complete annotated dataset is available and described in more details in Buche et al. (2021).

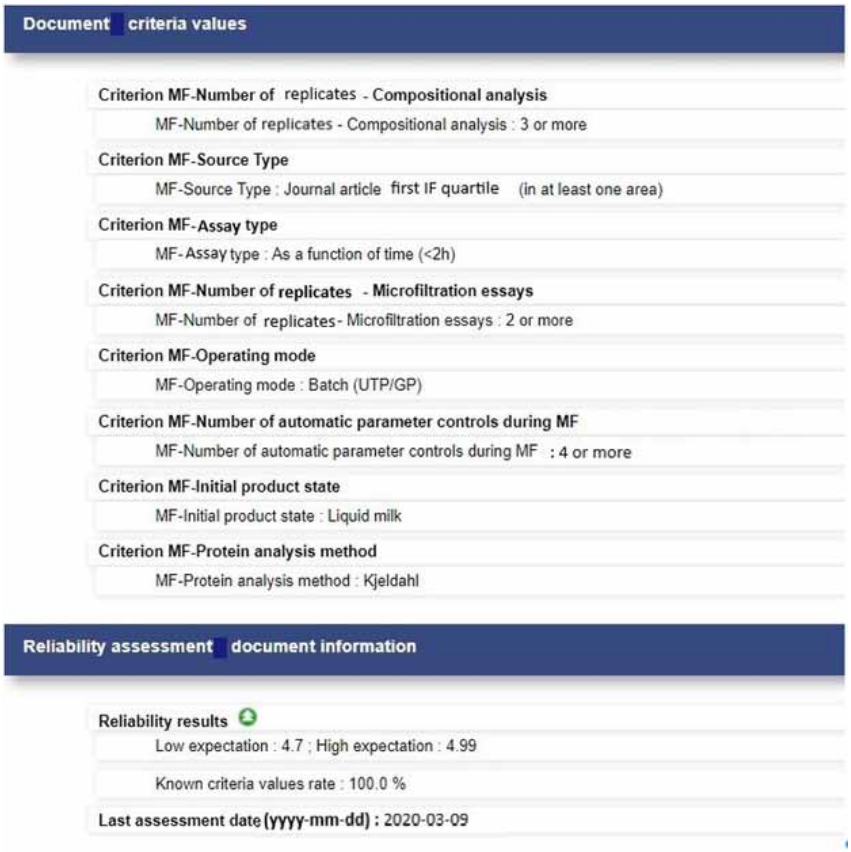
To assess the reliability of articles in the domain of milk microfiltration, experts were asked to provide a list of groups of metadata (Fig. 4 and the complete list in Buche et al. (2021). Each metadatum was associated with an expert opinion (see Buche et al. (2021)). The interval scores associated with each article were calculated based on the expert opinions and metadata values registered by annotators. For example, a list of metadata is associated with data from Jorgensen et al. (2016) (see Fig. 5). The range of reliability scores $\left[\underline{E_o}, \underline{E_o}\right]$ calculated for this document ([4.70, 4.99]) indicates that it is considered highly reliable. Based on these scores, @Web querying module computes reliability rank associated with each document. The reliability rank is an integer which ranges from 1 (most reliable) to n (least reliable). Documents are ranked in decreasing reliability as the reliability rank increases (see explanations about the ranking method in Destercke et al. (2013)). Reliability meta-information and reliability scores for the set of 52 articles used in this paper are available in Buche et al. (2021).

During operation of the microfiltration unit, the principle of crossflow filtration is to apply the flow of product to be treated (skimmed milk in this study) tangentially to the membrane and apply a difference of pressure between the two sides of the membrane to allow separation. The difference of pressure that forces the fluid to pass through the membrane is called the transmembrane pressure (TMP). The tangential flow at the membrane surface creates a shearing effect at the membrane surface, which in turn reduces fouling, removes the build up of retained particles at the membrane surface and reduces the drop in permeate flux through the membrane (Jp). The tangential flow can be characterised by either crossflow velocity or shear stress. The volume reduction factor (VRF), which

Table 2. Excerpt of the annotated table Dynamics of controlled microfiltration parameters including data from Jorgensen et al. (2016). VRF: volume reduction factor, STD: standard deviation, TMP: transmembrane pressure, JP: permeate flux

Experiment no.	Process step no.	Time	Time_UNIT	Process phase	VRF_AVG	VRF_STD	VRF_UNIT	Temperature	Temperature_UNIT	TMP	TMP_UNIT	JP_AVG	JP_UNIT
1	5	90	min	Constant phase (constant VRF)	2.5	0.1	One	50.1	°C	75.6	kPa	58.9	$L.h^{-1}.m^{-2}$
1	5	95	min	Constant phase (constant VRF)	2.5	0.1	One	50.1	°C	76.0	kPa	58.9	$L.h^{-1}.m^{-2}$
1	5	100	min	Constant phase (constant VRF)	2.5	0.1	One	50.1	°C	76.9	kPa	58.9	$L.h^{-1}.m^{-2}$

Figure 4. @Web reliability assessment associated with data from experiments of Jorgensen et al. (2016). MF: Microfiltration, IF: impact factor, UTP: Uniform Transmembrane Pressure, GP: Graded Permeability



equals the ratio of feed flow rate to retentate extraction flow rate, is used in the sector to obtain the targeted concentration of retained compounds in the final retentate.

Relationships among the operating parameters (i.e. VRF, TMP; Jp, shear stress, and crossflow velocity) depend on the configuration and membrane system used. Guided by the ontology of milk microfiltration, Fig. 6 displays the structure of the BN based on expert knowledge. The model considers three kinds of membranes: Tubular-UTP, Tubular GP-Isoflux and Tubular Classic (i.e. Tubular membranes with no UTP system, or GP®-Isoflux® membranes). VRF, shear stress, crossflow, Jp and TMP are discretized into 6, 5, 5, 9 and 13 categories, respectively. The categories were created based on the values usually expected for the type of membrane (see categories in Fig. 6).

The following parameters were estimated from three sources of knowledge (i.e. expert opinion, journal articles and experiments):

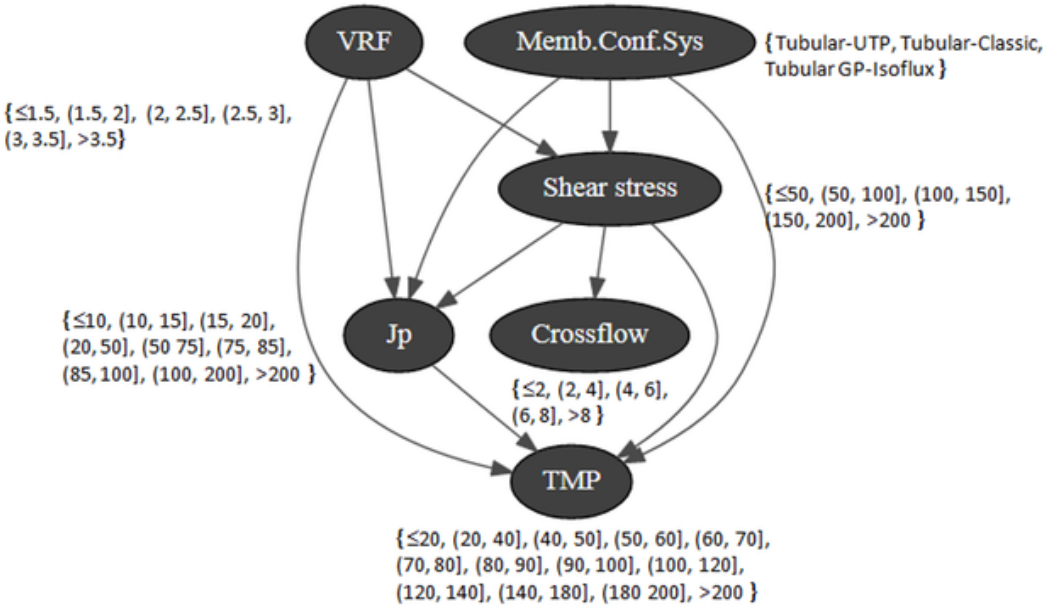
$$\theta_1 = P\left(VRF\right)$$

$$\theta_2 = P\left(Memb.Conf.Sys\right)$$

$$\theta_3 = P(Shear\,stress\,|VRF, Memb.Conf.Sys)$$

$$\theta_4 = P(Jp\,|VRF, Memb.Conf.Sys, Shear\,stress)$$

Figure 5. Bayesian network and the values of each variable modelling interactions in the network that occur in milk microfiltration using ceramic membranes. VRF: Volume reduction flux, Memb.Conf.Sys: configuration of the membrane system, Jp: Permeation flux, TMP: transmembrane pressure



$$\theta_5 = P(Crossflow \mid Shear\,stress)$$

$$\theta_6 = P(TMP \mid VRF, Memb.Conf.Sys, Jp, Shear\,Stress)$$

All parameters θ_i are initialised by expert opinion corresponding to the α_{ij} in Eq. 3. Based on experience, experts are able to explain part of the complex behaviour of the microfiltration process that they oversee. Table 3 displays an excerpt of the expert opinion regarding the shear stress given VRF and the kind of membrane (i.e. $\alpha_{3\{\leq 1.5, Tubular-UTP\}} = \{0.05, 0.15, 0.35, 0.35, 0.15\}$)

All parameters θ_i are then updated by means of Eq. 3 by using the learning dataset built from the extracted dataset and experiments according to the reliability rank of microfiltration articles and experiments. The dataset to which reliability ranks were assigned is composed of 893 experimental

Table 3. Conditional probability distribution P(Shear stress| volume reduction factor (VRF) = '≤1.5', membrane configuration systems (Memb.Conf.Sys) = "Tubular-UTP") extracted from the conditional probability distribution P(Shear stress| VRF, Memb. conf.sys) provided by expert opinion

VRF	Memb.Conf. Sys	Shear stress				
		≤50	(50, 100]	(100, 150]	(150, 200]	>200
...	...	...	...	...	...	...
≤1.5	Tubular-UTP	0.05	0.15	0.35	0.35	0.15
...	...	...	...	...	...	...

data points extracted from the RDF database using SPARQL queries from microfiltration articles, enriched with 33 confidential experiments. The reliability ranks range from 1 ("very reliable") to 16 ("not at all reliable"). A simple way to capture these reliability ranks in the learning dataset is to duplicate a datum (17-[reliability rank]) times (e.g., duplicate "very reliable" data 16 times). This approach has been applied to the extract of the learning dataset based on reliability ranks whose excerpt is presented in Table 4: The first (resp. third) line is duplicated 16 times (resp. 4 times) as the second one is not duplicated. The entire dataset with the experimental data from the literature is available on the INRAE dataverse (MICROFILTRATION learning dataset.

The model built and implemented using the python pyAgrum library based on the C++ aGrUM library (Gonzales *et al*., 2017) The model makes it possible to give two types of information through the estimation of probability distributions for:

1. TMP or Jp, given shear stress and VRF constraints and the type of membrane used (*i.e.* $P\left(Jp \mid VRF,\sigma_w,Memb.Conf.Sys\right)$ and $P\left(TMP \mid VRF,\sigma_w,Memb.Conf.Sys\right)$, respectively). For example, assuming that a Tubular-UTP membrane is associated with a VRF of (1.5, 2.0] and a shear stress of (50,100] Pa (Fig. 6a), the probability is 39% that Jp lies in the range (20,50] kg.h^{-1}.m^{-2} and 34% that TMP lies in the range (0,20] Pa.
2. Shear stress and VRF associated with the type of membrane used, given the expected TMP and Jp (*i.e* $P\left(VRF \mid Jp,TMP\right)$ $P\left(\sigma_w \mid Jp,TMP\right)$ and $P\left(Memb.Conf.Sys \mid Jp,TMP\right)$). This indicates that the model can estimate the values of control parameters which are most likely to meet the expected target (TMP, Jp) via $P\left(VRF,\sigma_w,Memb.Conf.Sys \mid \left(Jp,TMP\right)\right)$. For example, a tubular-UTP membrane must be used or $_w \in \left(50,100\right]$ and VRF∈(1.5,2.0] must be applied to ensure that TMP remains less than 20 Pa and Jp lies in the range (20, 50] kg.h^{-1}.m^{-2} (Fig. 6b).

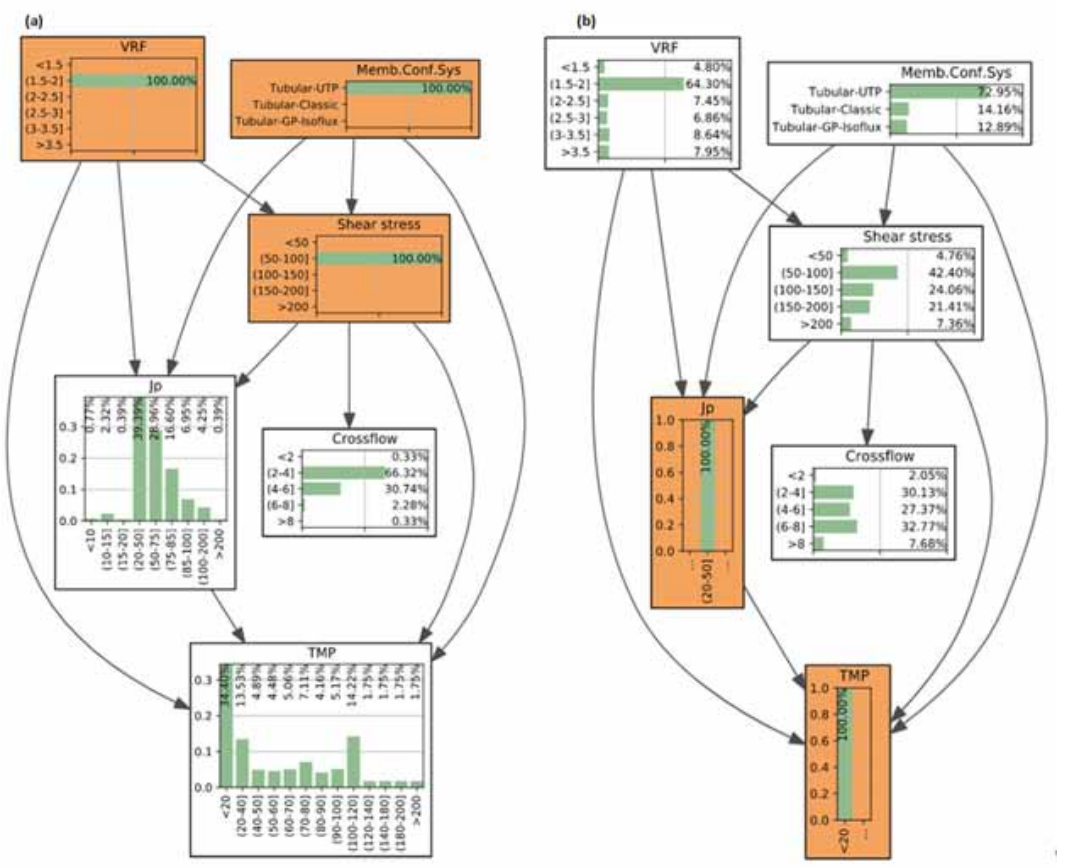
Validation of the model's predictive accuracy is based on leave-one-out cross-validation (Vehtari *et al*., 2017) and concerns the prediction of Jp and TMP given shear stress and VRF constraints and the type of membrane used. The leave-one-out cross-validation is performed in four configurations: complete or incomplete input data that include or do not include data reliability in parameter learning. For example, when considering missing data with data reliability, the validation yields a confusion matrix (i.e. predicted vs. raw data) for Jp (Table 5). Each one can be used to estimate the percentage of (1) predicted data that is actually raw data (i.e. *precision*) and (2) raw data that is predicted accurately (i.e. *recall*).

For example, the *precision* for Jp of 50-75 kg.h^{-1}.m^{-2} is the number of Jp correctly predicted in this range out of all Jp predicted in this range: 274/551=49.7%. Thus, 49.7% of the Jp that the model predicts as 50-75 kg.h^{-1}.m^{-2} is actually in this range (Table 5). In comparison, the *recall* for Jp of 50-75 kg.h^{-1}.m^{-2} is the number of Jp correctly predicted in this range out of the number of raw Jp in this range: 274/323=84.8%. Thus, the model predicts that 84.8% of the Jp of 50-75 kg.h^{-1}.m^{-2} is in this range. From matrix confusions, the overall accuracy of the model may be estimated at about 55% for Jp and TMP when the learning dataset contains missing data, regardless of the reliability of

Table 4. Extract of the learning dataset according to reliability rank. TMP: transmembrane pressure, VRF: volume reduction factor, Jp: permeation flux

Reliability rank	TMP (bar)	Shear stress (Pa)	VRF (-)	Membrane configuration	Jp (kg.h^{-1}.m^{-2})
1	3	162.7	1	Tubular-UTP	29.1
16	19.81	120	2	Tubular-UTP	50
13	500	Unknown	1	Tubular-Classic	70.36

Figure 6. (a) Estimated probability distributions of the transmembrane pressure (TMP) and permeation flux (Jp) (green bars in white boxes) given the evidence (green bars in orange boxes). (b) Estimated probability distributions of the shear stress and volume reduction factor (VRF) constraints (white boxes) given the Jp and TMP (orange boxes).



data (Table 6. Including data reliability increases the overall accuracy from 42% to 70% when the learning dataset has complete data (Table 6). However, the overall model accuracy for TMP is low (27-28%), due mainly to the existence of a limiting value of Jp for which a wide range of TMP is possible (Fig. 4.13 p.126 in Cheyran (1998)). As the model learned that nearly all TMP are possible when Jp is high and near its limiting value, it cannot discriminate TMP at high values of Jp. The solution for this issue is to consider the limiting Jp in order to determine the lowest TMP that can reach it. Increasing TMP beyond this lowest value is not useful, as the Jp can no longer increase. Using the dataset distribution to identify the limiting Jp is difficult because the latter is a function of the operating conditions of microfiltration (i.e. temperature, VRF, shear stress). This complexity helps explain why modelling skimmed milk microfiltration remains complex.

CONCLUSION AND PERSPECTIVES

This study developed a new practical and versatile workflow as a major step in building a decision support system for the agri-food sector. This new workflow combines knowledge and data integration guided by ontology and BNs. Based on expertise, the ontology paradigm is used to build the structure of the BN. The parameters of the model developed were estimated with integrated data

Table 5. Confusion matrix (i.e. predicted vs. raw data) for permeation flux (Jp) resulting from the leave-one-out cross-validation without including data reliability

		Predicted permeation flux data								
	"Value" (occurrence)	≤10 (0)	(10-15] (8)	(15-20] (0)	(20-50] (162)	(50-75] (551)	(75-85] (43)	(85-100] (10)	(100-200] (63)	>200 (38)
Raw permeation flux data	**≤10** (3)	0	0	0	0	3	0	0	0	0
	(10-15] (22)	0	5	0	2	15	0	0	0	0
	(15-20] (5)	0	1	0	2	2	0	0	0	0
	(20-50] (258)	0	2	0	124	127	5	0	0	0
	(50-75] (323)	0	0	0	25	274	20	4	0	0
	(75-85] (71)	0	0	0	4	54	12	0	1	0
	(85-100] (61)	0	0	0	2	54	2	1	2	0
	(100-200] (80)	0	0	0	3	22	4	4	42	5
	>200 (52)	0	0	0	0	0	0	1	18	33

Table 6. Overall model accuracy for permeation flux (Jp) and transmembrane pressure (TMP) according to the four configurations of parameter learning. Bold text indicates the highest accuracy for each parameter

Variable	Overall accuracy			
	Without data reliability and with missing data	With data reliability and missing data	Without data reliability and with complete data	With data reliability and complete data
Jp	57%	56%	42%	**70%**
TMP	**54%**	53%	28%	27%

from heterogeneous literature sources guided by a domain ontology. The reliability of the sources was included in the learning parameters via reliability indicators. The workflow was assessed for a complex food engineering process: skimmed milk microfiltration. The workflow can be used iteratively in order to enrich it regularly with new data and knowledge to enhance its deduction ability. An advantage of the approach is that the database connected to the model can be updated in an iterative process without damaging the entire system.

This article presents a complete initial iteration for skimmed milk microfiltration. A domain ontology of milk microfiltration was developed to annotate state-of-the-art literature sources, and a list of metadata was developed to assess data source reliability. They were used to build a structured database which was used in part to learn the BN. The Bayesian model was built to predict Jp and TMP process parameters as a function of the membrane technology and associated operating conditions, which is one innovation of this approach. The workflow iteration provided two results: (1) greater model accuracy for Jp when including the reliability of sources and (2) identification of knowledge gaps. The lower accuracy for TMP shows that the current model is not completely adapted to the physical phenomenon. This is due to two phenomena: (1) the ability to have several TMPs for a limiting

Jp and (2) the use of a uniform distribution when faced with a lack of knowledge, thus introducing a misleading uniformity (Walley, 1991). However the first results obtained with the proposed workflow based on ontology and BN are promising and must be extended.

In the iterative learning process, the presence of uniform distributions indicates the need for new experiments to provide more information to reduce entropy in the BN and bridge knowledge gaps. Another approach to address ignorance or potential uniformity is to use credal networks (Baudrit et al. 2016), which are an extension of BNs as a safer option to model imprecise information using convex sets of probabilities. To better describe the observed "plateau" effect of the TMPs with the limiting Jp, dynamic BNs (Baudrit et al., 2015) could be useful for predicting TMP as soon as Jp no longer changes.

The database will also be used to assess impacts of process parameters on milk quality parameters, which is another major aspect of process optimisation. A future perspective of this study is to demonstrate that the workflow developed can be adapted easily to predict other performance criteria of microfiltration (including protein transmission) and more generally other processes in the agri-food sector.

ACKNOWLEDGMENT

This study was supported by a grant from the Brittany Region (contract no. 16006734, INRA convention 30 0 01292), from FEDER (contract no. EU0 0 0171, INRA convention 30 0 01293) and ANR Datasusfood (ANR-19-DATA-0016).

REFERENCES

Astudillo-Castro, C., Cordova, A., Oyanedel-Craver, V., Soto-Maldonado, C., Valencia, P., Henriquez, P., & Jimenez-Flores, R. (2020). Prediction of the limiting flux and its correlation with the Reynolds number during the microfiltration of skim milk using an improved model. *Foods*, *9*(11), 1621. doi:10.3390/foods9111621 PMID:33172214

Astudillo-Castro, C. L. (2015). Limiting flux and critical transmembrane pressure determination using an exponential model: The effect of concentration factor, temperature, and cross-flow velocity during casein micelle concentration by microfiltration. *Industrial & Engineering Chemistry Research*, *54*(1), 414–425. doi:10.1021/ie5033292

AtWeb. (2021). https://www6.inrae.fr/cati-icat-atweb/

Aubin, S., Bisquert, P., Buche, P., Dibie-Barthelemy, J., Ibanescu, L., Jonquet, C., & Roussey, C. (2019). Recent progresses in data and knowledge integration for decision support in agri-food chains. *IC 2019- Journées francophones d'Ingénierie des Connaissances, AFIA*, 43-59. https://hal.archives-ouvertes.fr/hal-02284538

Barnabe, M., Blanc, N., Chabin, T., Delenne, J. Y., Duri, A., Frank, X., & Perrot, N. et al. (2018). Multiscale modeling for bioresources and bioproducts. *Innovative Food Science & Emerging Technologies*, *46*, 41–53. doi:10.1016/j.ifset.2017.09.015

Baudrit, C., Destercke, S., & Wuillemin, P. H. (2016). Unifying parameter learning and modelling complex systems with epistemic uncertainty using probability interval. *Information Sciences*, *367*, 630–647. doi:10.1016/j.ins.2016.07.003

Baudrit, C., Perrot, N., Brousset, J. M., Abbal, P., Guillemin, H., Perret, B., Goulet, E., Guerin, L., Barbeau, G., & Picque, D. (2015). A probabilistic graphical model for describing the grape berry maturity. *Computers and Electronics in Agriculture*, *118*, 124–135. doi:10.1016/j.compag.2015.08.019

Baudrit, C., Wuillemin, P. H., & Perrot, N. (2013). Parameter elicitation in probabilistic graphical models for modelling multi-scale food complex systems. *Journal of Food Engineering*, *115*(1), 1–10. doi:10.1016/j.jfoodeng.2012.09.012

Belna, M., Ndiaye, A., Taillandier, F., Agabriel, L., Marie, A. L., & Gésan-Guiziou, G. (2020). Formulating multiobjective optimization of 0.1 µm microfiltration of skim milk. *Food and Bioproducts Processing*, *124*, 244–257. doi:10.1016/j.fbp.2020.09.002

Brans, G. B. P. W., Schroën, C. G. P. H., Van der Sman, R. G. M., & Boom, R. M. (2004). Membrane fractionation of milk: State of the art and challenges. *Journal of Membrane Science*, *243*(1-2), 263–272. doi:10.1016/j.memsci.2004.06.029

Buche, P., Dervaux, S., Leconte, N., Belna, M., Granger-Delacroix, M., Garnier-Lambrouin, F., Gregory, G., Barrois, L., & Gesan-Guiziou, G. (2021). Milk microfiltration process dataset annotated from a collection of scientific papers. *Data in Brief*, *36*, 107063. doi:10.1016/j.dib.2021.107063 PMID:34026967

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., & Soler, L. (2013). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 805–819. doi:10.1109/TKDE.2011.245

Buron, M., Goasdoué, F., Manolescu, I., & Mugnier, M.-L. (2020). Ontology-Based RDF Integration of Heterogeneous Data. *EDBT*, *2020*, 299–310.

Carvalho, R. N., Laskey, K. B., & Costa, P. C. (2017). PR-OWL–a language for defining probabilistic ontologies. *International Journal of Approximate Reasoning*, *91*, 56–79. doi:10.1016/j.ijar.2017.08.011

Chapman, R., Cook, S., Donough, C., Lim, Y. L., Ho, P. V. V., Lo, K. W., & Oberthür, T. (2018). Using Bayesian networks to predict future yield functions with data from commercial oil palm plantations: A proof of concept analysis. *Computers and Electronics in Agriculture*, *151*, 338–348. doi:10.1016/j.compag.2018.06.006

Cheryan, M. (1998). *Ultrafiltration and microfiltration handbook*. CRC Press. doi:10.1201/9781482278743

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*(2-3), 393–405. doi:10.1016/0004-3702(90)90060-D

Destercke, S., Buche, P., & Charnomordic, B. (2013). Evaluating Data Reliability: An Evidential Answer with Application to a Web-Enabled Data Warehouse. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 92–105. doi:10.1109/TKDE.2011.179

Devitt, A., Danev, B., & Matusikova, K. (2006). *Constructing Bayesian networks automatically using ontologies*. Academic Press.

Drury, B., Valverde-Rebaza, J., Moura, M. F., & de Andrade Lopes, A. (2017). A survey of the applications of Bayesian networks in agriculture. *Engineering Applications of Artificial Intelligence*, *65*, 29–42. doi:10.1016/j.engappai.2017.07.003

Fenz, S. (2012). An ontology-based approach for constructing Bayesian networks. *Data & Knowledge Engineering*, *73*, 73–88. doi:10.1016/j.datak.2011.12.001

Gésan-Guiziou, G. (2010). Separation technologies in dairy and egg processing Part II. Separation technologies in the processing of particular foods and nutraceuticals. In Separation, Extraction and Concentration Processes in the Food, Beverage and Nutraceutical Industries (pp. 341–380). Woodhead Publishing Limited. doi:10.1533/9780857090751.2.341

Gésan-Guiziou, G., Boyaval, E., & Daufin, G. (1999). Critical stability conditions in crossflow microfiltration of skimmed milk: Transition to irreversible deposition. *Journal of Membrane Science*, *158*(1-2), 211–222. doi:10.1016/S0376-7388(99)00017-4

Gonzales, C., Torti, L., & Wuillemin, P. (2017). agrum: a graphical universal modem framework. In *Proceedings of the 30th International Conference on Industrial, Engineering, Other Applications of Applied Intelligent Systems*. Springer-Verlag.

Guillard, G., Couvert, C., Stahl, V., Buche, P., Hanin, A., Denis, C., Dibie, J., Dervaux, S., Loriot, C., Vincelot, T., Huchet, V., Perret, B., & Thuault, D. (2017). MAP-OPT: A software for supporting decision-making in the field of modified atmosphere packaging of fresh non respiring foods. Packaging Research. *De Gruyter OPEN*, *2*(1), 28–47. doi:10.1515/pacres-2017-0004

Heckerman, D. (2008). *A Tutorial on Learning with Bayesian Networks*. Springer. doi:10.1007/978-3-540-85066-3_3

Ishak, M. B., Leray, P., & Amor, N. B. (2011). Ontology-based generation of object oriented bayesian networks. In BMAW 2011 (pp. 9-17). Academic Press.

Jalilvand, Z., Ashtiani, F. Z., Fouladitajar, A., & Rezaei, H. (2014). Computational fluid dynamics modeling and experimental study of continuous and pulsatile flow in flat sheet microfiltration membranes. *Journal of Membrane Science*, *450*, 207–214. doi:10.1016/j.memsci.2013.09.008

Jensen, F. V., & Nielsen, T. D. (2007). Bayesian Networks and Decision Graphs. Springer Verlag. doi:10.1007/978-0-387-68282-2

Ji, Z., Xia, Q., & Meng, G. (2015, August). A review of parameter learning methods in Bayesian network. In *International Conference on Intelligent Computing* (pp. 3-12). Springer. doi:10.1007/978-3-319-22053-6_1

Jorgensen, C. E., Abrahamsen, R. K., Rukke, E.-O., Johansen, A.-G., Schüller, R. B., & Skeie, S. B. (2016). Optimization of protein fractionation by skim milk microfiltration: Choice of ceramic membrane pore size and filtration temperature. *Journal of Dairy Science*, *99*(8), 6164–6179. doi:10.3168/jds.2016-11090 PMID:27265169

Laskey, K. B. (2008). MEBN: A language for first-order Bayesian knowledge bases. *Artificial Intelligence*, *172*(2-3), 140–178. doi:10.1016/j.artint.2007.09.006

Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., & Vialle, C. (2016). Decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Computers and Electronics in Agriculture*, *127*, 351–367. doi:10.1016/j.compag.2016.06.020

MICROFILTRATION learning dataset. (n.d.). https://doi.org/10.15454/DYNCEN

MICROFILTRATION ontology. (n.d.). http://agroportal.lirmm.fr/ontologies/MICROFILTRATION

Munch, M., Buche, P., Manfredotti, C., Wuillemin, P. H., & Angellier-Coussy, H. (2021). A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models: application to processing of bio-composites for food packaging. *15th International Conference on Metadata and Semantics Research.*

Munch, M., Dibie-Barthelemy, J., Wuillemin, P. H., & Manfredotti, C. (2019, May). Towards interactive causal relation discovery driven by an ontology. FLAIRS 32.

Munch, M., Wuillemin, P. H., Manfredotti, C., Dibie, J., & Dervaux, S. (2017). Learning probabilistic relational models using an ontology of transformation processes. In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 198-215). Springer.

O'Hagan, A. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. Wiley.

Pan, R., Ding, Z., Yu, Y., & Peng, Y. (2005). A Bayesian network approach to ontology mapping. In *International Semantic Web Conference* (pp. 563-577). Springer.

Patnaikuni, P., Shrinivasan, R., & Gengaje, S. R. (2017). Survey of Multi Entity Bayesian Networks (MEBN) and its applications in probabilistic reasoning. *International Journal of Advanced Research in Computer Science*, *8*(5), 2425–2429.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Perrot, N., De Vries, H., Lutton, E., Van Mil, H. G., Donner, M., Tonda, A., & Axelos, M. A. et al. (2016). Some remarks on computational approaches towards sustainable complex agri-food systems. *Trends in Food Science & Technology*, *48*, 88–101.

Saboya, L. V., & Maubois, J. L. (2000). Current developments of microfiltration technology in the dairy industry. *Le Lait*, *80*(6), 541–553.

Salmerón, A., Rumí, R., Langseth, H., Nielsen, T. D., & Madsen, A. L. (2018). A review of inference algorithms for hybrid Bayesian networks. *Journal of Artificial Intelligence Research*, *62*, 799–828.

Sandblom, R. M. (1974). *Filtering process.* Swedish Patent, 7416257.

Setiawan, F. A., Budiardjo, E. K., Basaruddin, T., & Aminah, S. (2017, December). A Systematic Literature Review on Combining Ontology with Bayesian Network to Support Logical and Probabilistic Reasoning. In *Proceedings of the 2017 International Conference on Software and e-Business* (pp. 1-12). Academic Press.

Skrzypek, M., & Burger, M. (2010). Isoflux® ceramic membranes—Practical experiences in dairy industry. *Desalination*, *250*(3), 1095–1100. doi:10.1016/j.desal.2009.09.116

Staab, S., & Studer, R. (Eds.). (2009). Handbook on ontologies (2nd. ed.). Springer Publishing Company, Incorporated, ISBN 978-3-540-70999-2, 811pp.

Truong, B.A., Lee, Y., & Lee, S.Y. (2005). A Unified Context Model: Bringing Probabilistic Models to Context Ontology. In: Enokido, T., Yan, L., Xiao, B., Kim, D., Dai, Y., Yang, L.T. (eds) Embedded and Ubiquitous Computing – EUC 2005 Workshops. EUC 2005. Lecture Notes in Computer Science, vol 3823. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11596042_59

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, *11*(2), 93–136. doi:10.1017/S0269888900007797

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. doi:10.1007/s11222-016-9696-4

Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall. doi: 10.2307/2347427

Yang Y., & Calmet, J. (2005). OntoBayes: An Ontology-Driven Uncertainty Model. IEEE International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06), 2005, pp. 457-463, doi: 10.1109/CIMCA.2005.1631307.

Yeumo, D. E., Alaux, M., & Arnaud, E. (2017). Developing data interoperability using standards: A wheat community use case [version 2; peer review: 2 approved]. *F1000Research, 6,* 1843. https://doi.org/10.12688/f1000research.12234.2

Yun, B., Buche, P., Bisquert, P., Costa, S., & Croitoru, M. (2018). Consumer perception data and scientific arguments about food packaging functionalities for fresh strawberries. *Data in Brief, Elsevier*, *20*(October), 1924–1927. doi:10.1016/j.dib.2018.09.034

APPENDIX

An example of RDF data corresponding to the content of the cell associated with column Time for the first line of data presented in Table 2, is given below:

```
<onto:hasForCell>
<onto:Cell rdf:about=”Cell-14_Row-23_4150”>
<rdf:type rdf:resource=”/resources/MICROFILTRATION#time”/>
<onto:hasForOriginalValue>Time</onto:hasForOriginalValue>
<onto:hasForColumnNumber rdf:datatype=”http://www.w3.org/2001/XMLSchema#integer”
>4</onto:hasForColumnNumber>
<onto:hasForFS>
<onto:CFS rdf:about=”CFS_CELL-14_Row-23_4150”>
<rdf:type rdf:resource=”/resources/atWeb/annotation/Scalar”/>
<onto:hasForUnit rdf:resource=”/resources/MICROFILTRATION#Minute”/>
<onto:hasForFuzzyElement>
<onto:FuzzySet rdf:about=”FS_Cell-14_Row-23_4150”>
<onto:hasForMaxKernel>90</onto:hasForMaxKernel>
<onto:hasForMinKernel>90</onto:hasForMinKernel>
<onto:hasForMinSupport>90</onto:hasForMinSupport>
<onto:hasForMaxSupport>90</onto:hasForMaxSupport>
</onto:FuzzySet>
</onto:hasForFuzzyElement>
</onto:CFS>
</onto:hasForFS>
</onto:Cell>
```

An example of SPARQL query to retrieve the subset of columns (Experience_number, ProcessStep_number, Time, VRF) presented in Table 2, arguments of the relation microfiltration_controlled_parameter_evolution_relation is given below:

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix xsd: <http://www.w3.org/2001/XMLSchema#>
prefix atweb: <http://opendata.inra.fr/resources/MICROFILTRATION#>
prefix atweb-data: <https://opendata.inra.fr/resources/atWeb/annotation/>
prefix atweb-core: <http://opendata.inra.fr/resources/core#>
SELECT ?minKernel_e_n ?minKernel_p_s_n ?minKernel_t ?unit_t ?minKernel_vrf
WHERE {
# Relation Microfiltration_controlled_parameter_evolution_relation is queried
?Relation rdf:type atweb:microfiltration_controlled_parameter_evolution_relation.
?Relation atweb-core:hasAccessConcept ?accessConceptExperience_number.
?accessConceptExperience_number rdf:type atweb:experience_number .
?accessConceptExperience_number atweb-data:hasForFS ?fuzzySet_e_n.
?fuzzySet_e_n atweb-data:hasForUnit ?unit_e_n.
?fuzzySet_e_n atweb-data:hasForFuzzyElement ?fuzzyElement_e_n.
?fuzzyElement_e_n atweb-data:hasForMinKernel ?minKernel_e_n .
?fuzzyElement_e_n atweb-data:hasForMaxKernel ?maxKernel_e_n.
```

?Relation atweb-core:hasAccessConcept ?accessConceptProcessStep_number.
?accessConceptProcessStep_number rdf:type atweb:process_step_number .
?accessConceptProcessStep_number atweb-data:hasForFS ?fuzzySet_p_s_n.
?fuzzySet_p_s_n atweb-data:hasForUnit ?unit_p_s_n.
?fuzzySet_p_s_n atweb-data:hasForFuzzyElement ?fuzzyElement_p_s_n.
?fuzzyElement_p_s_n atweb-data:hasForMinKernel ?minKernel_p_s_n .
?fuzzyElement_p_s_n atweb-data:hasForMaxKernel ?maxKernel_p_s_n.
?Relation atweb-core:hasAccessConcept ?accessTime.
?accessTime rdf:type atweb:time .
?accessTime atweb-data:hasForFS ?fuzzySet_t.
?fuzzySet_t atweb-data:hasForUnit ?unit_t.
?fuzzySet_t atweb-data:hasForFuzzyElement ?fuzzyElement_t.
?fuzzyElement_t atweb-data:hasForMinKernel ?minKernel_t .
?fuzzyElement_t atweb-data:hasForMaxKernel ?maxKernel_t .
?Relation atweb-core:hasAccessConcept ?accessVRF.
?accessVRF rdf:type atweb:vrf .
?accessVRF atweb-data:hasForFS ?fuzzySet_vrf.
?fuzzySet_vrf atweb-data:hasForUnit ?unit_vrf.
?fuzzySet_vrf atweb-data:hasForFuzzyElement ?fuzzyElement_vrf.
?fuzzyElement_vrf atweb-data:hasForMinKernel ?minKernel_vrf .
?fuzzyElement_vrf atweb-data:hasForMaxKernel ?maxKernel_vrf .
}

Cédric Baudrit is a researcher for the National Research Institute for Agriculture, Food and the Environment (INRAE), in knowledge representation and reasoning. He is interested in the development of mathematical tools capable of (1) integrating fragmented heterogeneous knowledge stemming from different sources; (2) taking into account stochastic and epistemic uncertainty in order to model global complex system. In 2005, he received the Ph.D. degree in computer science from Université Paul Sabatier, in Toulouse (France). In 2002, he holds a master in Applied Mathematics and Computer Science from the University of Orléans, France.

Patrice Buche received the PhD degree in computer science from the University of Rennes, France, in 1990. He has been assistant professor with AgroParisTech, Paris from 1992 to 2002 and a research engineer with INRAE, Agricultural Research Institute since 2002. His research works mainly concern data and knowledge integration from heterogeneous sources, fuzzy querying in structured and weakly structured databases and argumentation, with applied projects in food and biobased product engineering.

Christophe Fernandez is a software developer in a team of two searchers specialized both mathematics applied and artificial intelligence. He has developed software dedicated to knowledge transfer or multi-objective optimization under INRAE copyrights. He is also co-writer of about fifteen scientific articles.