Combining ontology and probabilistic models for the design of bio-based product transformation processes

Mélanie Munch, Patrice Buche, Stéphane Dervaux, Juliette Dibie, Liliana L. Ibanescu, Cristina Manfredotti, Pierre-Henri Wuillemin, Helene Angellier-Coussy

▶ **To cite this version:**

Mélanie Munch, Patrice Buche, Stéphane Dervaux, Juliette Dibie, Liliana L. Ibanescu, et al.. Combining ontology and probabilistic models for the design of bio-based product transformation processes. Expert Systems with Applications, 2022, 203, pp.117406. 10.1016/j.eswa.2022.117406 . hal-03662183

HAL Id: hal-03662183
https://hal.science/hal-03662183

Submitted on 15 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining Ontology and Probabilistic Models for the Design of Bio-based Product Transformation Processes

Mélanie Munch[a,*], Patrice Buche[a,d], Stéphane Dervaux[b], Juliette Dibie[b], Liliana Ibanescu[b], Cristina Manfredotti[b], Pierre-Henri Wuillemin[c], Hélène Angellier-Coussy[a]

[a]*IATE, Univ Montpellier, INRAE, Institut Agro, Montpellier, F-34060, France*
[b]*UMR MIA-Paris, AgroParisTech, INRAE, Université Paris-Saclay, Paris, F-75005, France*
[c]*Sorbonne Universites, UPMC, U. Paris 06, CNRS UMR, LIP6, Paris, F-75005, France*
[d]*LIRMM, U. Montpellier, CNRS, INRIA GraphIK, Montpellier, F-34060, France*

*Corresponding author

Email addresses: melanie.munch@gmail.com (Mélanie Munch), patrice.buche@inrae.fr (Patrice Buche), stephane.dervaux@inrae.fr (Stéphane Dervaux), juliette.dibie-barthelemy@inrae.fr (Juliette Dibie), liliana.ibanescu@agroparistech.fr (Liliana Ibanescu), cristina.manfredotti@agroparistech.fr (Cristina Manfredotti), pierre-henri.wuillemin@lip6.fr (Pierre-Henri Wuillemin), helene.coussy@umontpellier.fr (Hélène Angellier-Coussy)

Preprint submitted to Expert Systems with Applications *June 15, 2022*

Abstract

This paper presents a workflow for the design of transformation processes using different kinds of expert's knowledge. It introduces POND (Process and observation ONtology Discovery), a workflow dedicated to answer expert's questions about processes. It addresses two main issues: 1) how to represent the processes inner complexity, and 2) how to reason about processes taking into account uncertainty and causality. First, we show how to use a semantic model, an ontology, and its associated data to answer some of the expert's questions concerning the processes, using semantic web languages and technologies. Then, we describe how to learn a predictive model, to discover new knowledge and provide explicative models by integrating the semantic model into a probabilistic relational model. The result is a complete workflow able to extensively analyse transformation processes through all their granularity levels and answer expert's questions about their domains. An example of this workflow is given on biocomposites manufacturing for food packaging.

Keywords: Ontologies, Probabilistic Relational Models, Knowledge Discovery, Causality

1. Introduction

In industry, a *production process* defines the *steps* through which raw materials (or *inputs*) are transformed into a final product (or *output*). The particularity -and difficulty- of its analysis lays on its heterogeneity (different variables, different scales) and time-dependent flow of information. Indeed, in order to analyse or to compare different processes, there is the need to represent the measurements (or *observations*) that estimate or calculate the value of a property (denoted as *characteristics* or *attribute*), associated either with an input/output, either with a step. In this paper we will use the term *observation* for a group of one or more *attributes*[1] having a result that may have a qualitative value or a quantitative value.

Analysing a process often requires to discover new knowledge. For instance, given a cooking recipe, it might be interesting to see whether modifying the quantity of sugar or the oven's temperature would have an influence

[1] Attributes will be denoted as *variables* if integrated in a probabilistic model.

over the final product. To answer this question, we would first need multiple repetitions of the recipe to have a good dataset able to represent the different situations. However, depending on external conditions, the precision of the different measures can vary: sometimes the measured quantity of sugar might be slightly off, or a hot weather might have an influence over the cake's cooking. These small variations, that characterize uncertainty, need to be addressed by the model. On another hand, defining the impact of a parameter over others can be a complex task, that requires interventions (Pearl, 2009) (i.e. assessing whether a variable is influenced by another by forcing some values). If those are sometimes easy to do (I can force the temperature of my oven), in some cases, they are impossible to assess (I can't force my cake's final taste to check whether the quantity of sugar is influenced). As a consequence, we want our model to be able to integrate causal knowledge in order to discover such relations without having to rely on interventions. For all these reasons, discovering new knowledge within a *transformation process* requires, on top of a robust system to represent and organize all of the concepts seen before, a reasoning model, able to deal with both (1) the uncertainty that stems from the variability of transformation processes and (2) the eventual causality needed for in-depth analysis.

In this article we present the workflow POND (Process and observation ONtology Discovery) **that provides tools to represent and reason about a transformation process in its overall complexity**.

To do so, POND needs to address two main challenges:

- The representation of the transformation processes and their inherent complexity and heterogeneity;
- The representation of uncertainty in coherent models able to answer specific (sometimes causal) questions about the domain.

In order to address the first challenge and to provide data and knowledge integration, a relevant solution is the use of an ontology (Doan et al., 2012). An ontology gives a structured and formalized representation of the specific vocabulary from a domain (Gruber, 2016; Guarino et al., 2009), using some logical language (e.g. Description Logic (DL) (Baader et al., 2017)) and allowing logical operation (e.g. inference for deducing new information). Lots of methods, technologies and tools exist for building and using ontologies, some of them becoming standards promoted either by the

OBO Foundry (OBO), a community development of interoperable ontologies for the biological sciences, or by the World Wide Web Consortium (W3C) (W3C), an international community that develops open standards to ensure the long-term growth of the Web. Moreover, publishing ontologies on the Linked Open Data (LOD) cloud (LOD) and building networks of interconnected ontologies (Suárez-Figueroa et al., 2012) should facilitate data integration and data sharing, such as giving access to data from specific disciplines or data produced within specific geographic regions (Bizer, 2013).

In order to address the second challenge, we propose the use of two probabilistic models, the Bayesian Network (BN) (Pearl, 1985) and the Probabilistic Relational Model (PRM) (Friedman et al., 1999), that allow to represent variables and their influence on each other. The main idea is to use the semantic knowledge encompassed in the ontologies to guide the learning of the probabilistic model, which has already been proposed in our previous works Munch et al. (2017, 2019a). The introduction of expert's knowledge allows us to introduce causal constraints which give a new overview on the data and a framework for causal discovery.

The goal of the POND workflow is to propose a way of representing and reasoning on data extracted from different sources about a specific transformation process. Our originality stems from (1) the adaptability of the representation part, that allows the combination of two knowledge sources (the ontology on one hand, and the expert's inputs on the other); and (2) the scope of the questions that can be answered through this workflow, some being answered by directly querying the data, and others by analysing a model, learned for the occasion, that is able to reason with the transformation process's complexity. In this article, we denote as *expert* a person (or group of persons) who possesses knowledge on a specific domain that can be represented in the Process and Observation Ontology (PO^2). Experts are considered at the same time as users of POND (they express their wish to reason about a transformation process and use POND at this purpose), and as providers of expert knowledge (necessary for the workflow). Their different interventions are summarized in Figure 2.

The paper is organized as follows. Section 2 presents the state of the art necessary to introduce both the ontologies and the PRMs. Section 3 describes the POND system and its architecture. It focuses on three parts: the PO^2 ontology, the mapping between the ontology and the PRM, and the exploitation of the PRM. The fourth section, Section 4, presents the application of the POND workflow for the design of biomass transformation

processes. Finally, Section 5 discusses the results and Section 6 concludes this paper.

2. State of the art

This article presents POND, a workflow dedicated to structure and exploit knowledge to reason about transformation processes. In this section we introduce the concepts of ontologies, BNs, PRMs, required to build our workflow. As we will see in more details in Section 3, ontologies can be used to address both the structuration issue as well as to answer complex or simple descriptive questions about the data. However, to deal with more complex queries (for instance those introducing uncertainty), in this section, we introduce as well an overview on the combination of ontology and PRM and on causal discovery.

2.1. Ontologies

In the context of computer and information sciences, *ontology* is the term used to refer to a shared understanding of some domain of interest and it is often conceived as a set of classes, attributes (or properties), and relationships (or relations among class members) (Gruber, 2016; Guarino et al., 2009). An ontology is usually specified in languages corresponding to first-order logic fragments, allowing abstraction from data structures and implementation, and providing reasoning capabilities. Therefore, ontologies are considered as semantic models and are used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces to independent, knowledge-based services.

There exists standard languages to define ontologies (e.g. Open Biomedical Ontology (OBO) format (OBO) and the Web Ontology Language (OWL) (OWL)), methodologies guiding how to build an ontology (e.g. NeOn methodology (Suárez-Figueroa et al., 2012) and Linked Open Terms (LOT) Methodology (Poveda-Villalón et al., 2019)), Ontology Designed Patterns (Gangemi & Presutti, 2009) as modeling solutions that address recurrent ontology design problems, repositories storing ontologies (e.g. BioPortal (Bio), OBO Foundry (OBO), Linked Open Vocabularies (LOV)) and a variety of commercial and open source tools for creating and using ontologies.

The first step in building an ontology is to define its requirements specification, i.e. to define its scope and its end-users and to generate a list of competency questions (Suárez-Figueroa et al., 2012; Poveda-Villalón et al.,

2019). When tackling the problem of the representation of a transformation process in food industry, the following requirements must be addressed: i) describe each step of the process by a set of experimental observations available for the inputs and outputs, ii) represent in a precise, quantitative manner the inputs and the outputs using some characteristics of interest (attributes), and iii) allow the comparison of different transformation processes in order to enhance product formulation.

In the OBO Foundry (OBO) there exists some ontologies that covers the subject of process modelling in biology: there are some generic ontologies, e.g. Information Artifact Ontology (IAO) (IAO), Relations ontology (RO) (RO) and Basic Formal Ontology (BFO) (BFO), and more specific ontologies, e.g. Ontology for Biomedical Investigations (OBI) (Bandrowski et al., 2016). The Core Ontology for Biology and Biomedicine (COB) (COB) is an under development ontology aiming to merge key classes and relations of the OBO Foundry ontologies.

The FoodOn Ontology (Dooley et al., 2018) aims to be an harmonized food ontology, that inherits the terms from the LanguaL Thesaurus (Ireland & Møller, 2010; Lan), a well known vocabulary and a system used to describe data about food (Ireland & Møller, 2000, 2016). Several working groups are restructuring FoodOn sub-hierarchies, one of them concerning 250 existing food transformation processes (Dooley et al., 2021).

The Semantic Sensor Network (SSN) (SSN) ontology is an ontology for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so and the observed properties, as well as actuators. It includes a lightweight but self-contained core ontology called SOSA (Sensor, Observation, Sample, and Actuator) (Janowicz et al., 2018; SOS). SSN and SOSA are the current recommendations promoted by the Open Geospatial Consortium (OGC) (OGC) and by the World Wide Web Consortium (W3C) (W3C).

A simple-to-use generic ontological model to represent transformation processes is well suited to manage data extraction from heterogeneous literature sources as shown in Lousteau-Cazalet et al. (2016). Moreover, representing relations between data in this ontological core model as n-ary relations Buche et al. (2013) facilitates the usage by end-users who are familiar with entering and manipulating data in spreadsheets. Yet, this simple model, based on a tabular representation, becomes insufficient when it is necessary to model complex transformation processes requiring to link output components of a step to input component of others. The Ontology for Food Process

Experiment (OFPE) Muljarto et al. (2014) proposes a generic semantic food transformation model based on OWL (OWL) which is built on four main classes: Product, Operation, Attribute and Observation. OFPE is extended by the PO^2 Ontology (Process and Observation Ontology) which allows us to represent a generic transformation process described by a set of experimental observations available for the inputs and outputs of each step of the production process. PO^2 is based on seven main classes: Component, Step, Attribute, Observation, Material, Method, Scale. The PO^2 version 1.5 published in (Ibanescu et al., 2016) evolved into the current 2.2.1 version (PO2, a), one of the contributions of this paper, described in Section 3.1.1. This version of the PO^2 core model reuses various existing ontologies (BFO (BFO), SOSA (Janowicz et al., 2018; SOS), IAO (IAO) and Time Ontology (TIM)) enhancing interoperability purposes.

A Food Process Modelling working group started two years ago is trying to compare the different existing process models and to provide a generalized process ontology. Among the investigated models there are OBI (Ontology for Biomedical Investigations) (Bandrowski et al., 2016), Exact2 (EXperimental ACTions) (Soldatova et al., 2014), Time Ontology (TIM), SOSA (Janowicz et al., 2018; SOS) and PO^2 (PO2, a). OBI and Exact2 are reusing BFO, each process having a role, while the time aspect is not considered. The focus is on categorizing the different processes used in biology according to the BFO and IAO main concepts. The main limit identified concerning the use of OBI or Exact2 as a generalized process ontology is that time is not integrated. SOSA (Janowicz et al., 2018; SOS) is a core ontology for the SSN ontology and allows us to represent sensors, observations, samples and actuators. The latest version of the PO^2 Ontology (PO2, a), contribution of this paper, is reusing concepts from SOSA, from the Time Ontology and from BFO.

PO^2 Ontology as a core ontology is compliant with a set of Competency Questions (see Section 3) listed by a group of experts in the field of bio-based product transformation processes in the framework of several national and international projects. For example, PO^2 core ontology is used in this paper to represent biocomposite manufacturing for food packaging experiments partially produced in H2020 ECOBIOCAP, NOAW and RESURBIS projects. A supplementary reason to choose the PO^2 core model is the development of dedicated tools to facilitate knowledge base implementation, another contribution of this paper: (i) PO^2 Manager (Dervaux et al., 2018; Buche et al., 2020), a standalone application, described in Section 3.1.2, that

assists domain experts in extending the PO^2 core model for a specific domain (food packaging making in this paper) and editing data using PO^2; (ii) SPO^2Q (Buche et al., 2020), a Web application described in Section 3.2 assisting users in querying data sets structured using the PO^2 core model.

2.2. Bayesian Networks and Probabilistic Relational Models

As mentioned before, our problem requires a model able to deal with uncertainty while offering a good integration of expert knowledge. To do so, we present in this section two probabilistic graphical models, the Bayesian Networks (BNs) (Pearl, 1985) and their object-oriented extension, the Probabilistic Relational Models (PRMs) (Friedman et al., 1999). While both allow to represent domains under the form of directed acyclic graphs, with variables as nodes and their relations encoded as arcs, PRMs offer one additional layer in order to better model expert's knowledge constraints.

A BN's graph G is defined as G=(V,E), where V and E are respectively the sets of all its nodes (representing discretized variables) and arcs (representing the conditional dependencies). To each variable, a conditional probability table is associated, giving the probability distribution for each possible value it can take and how the values of its parents (i.e. variables that have an oriented path toward that variable) influence it. Far from a black box system, this offers a useful double-reading: given X and Y two variables with an arc between them ($X \rightarrow Y$), both qualitative ("Has X an influence over Y?", i.e. "Is there a oriented path from X to Y in G?") and quantitative ("What is the influence of X over Y?", i.e. "What is the impact of X's value over Y's value according to the conditional probability table?") questions can be answered. Since arcs encode independence, a variable's value only depends on those of its parents, allowing to clearly see how variables are tied together.

PRMs also use this principle, but they add a layer of description in order to specify groups of variables and their relations. Indeed, on the contrary of BNs that are only defined on one level (the directed acyclic graph), the variables and their relations in a PRM are described on two levels (as illustrated in Figure 1 (a) and (b)):

- The **Relational Schema** (Figure 1 (a)) defines groups of attributes as classes, and relations between those classes. It is important to note that at this point, there are no relation between attributes: only classes are linked, through so called **relational slots**.

- The **Relational Model** (Figure 1 (b)) defines the relations between attributes. These can be intra-class (the two attributes are defined in the same class, e.g. X and Y in class A), or inter-classes (the attributes are attached to two different classes, e.g. Y and U). In the case of inter-classes relations, a constraint is given on its orientation, as it has to follow the orientation of the relational slot between the two classes. If there is no relational slot, then there cannot be a relation between the variables.

Similarly to BNs, the PRM's different level can be learned or manually built. However, an interesting feature is that once a relational schema is defined, the relational model can be learned in a similar way to a BN (Getoor & Taskar, 2007). In our work, we (1) integrate expert knowledge in the relational schema by manually defining classes of attributes and relational slots; then (2) automatically learn from data the relational model over the relational schema defined under these expert constraints, using classical Bayesian networks learning algorithm[2]. Once the relational model is defined, the classes can be instantiated in a system similar to a BN (Figure 1 (c)), allowing us to use tools dedicated to BN's analysis to reason on the process. As presented in the following sections, this distinction in two steps allows us to deduce causal relations from our learned model and compensate the possible insufficiency of data.

To conclude this part, we have however to point out that despite being well-equipped to represent correlations between variables, BN (and PRM)'s relations do not represent causation: given $X \rightarrow Y$, there is no direct reading that allows to tell whether X causes Y or the contrary. Yet, since in our case they are learned under expert's knowledge (that we suppose equivalent to causal constraints), then it becomes possible to deduce some causality by looking at the **Essential Graph** (EG) (Madigan et al., 1996), a semi-directed graph that can be deduced from any BN and which represents its Markov's equivalence class. The interest of the EG lies on the fact that, despite having the same structure as the BN, its arcs orientations reflect whether they depend on the data used to learn the model or not. If an arc is oriented both in the BN and its EG, then it means that it cannot be changed without contradicting the inherent independence of the learning dataset. On

[2]In this article, we use a classical Greedy Hill Climbing algorithm Chickering (2003) with the BIC score Schwarz (1978).

the contrary, if an arc is not oriented, then it means that its orientation can be reversed in the BN without contradicting these independences. Since we are working on a BN learned under causal constraints (thanks to the expert's knowledge), we suppose, in the following, that oriented arcs in the EG can be thought as causal. This supposes a set of specific conditions that we detail further in Section 2.4.

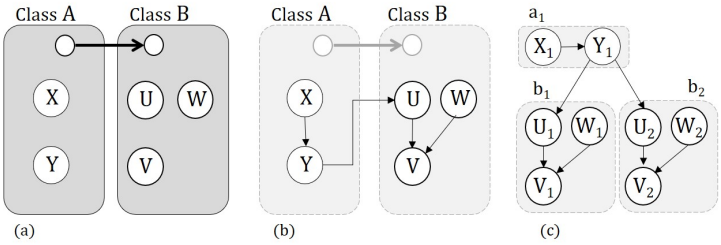

Figure 1: A PRM is described by two components: (a) the relational schema defines groups of attributes as classes and their relations; (b) the relational model gives the probabilistic dependencies between the attributes. Once defined, the classes can be instantiated: the attributes of each class in the relational model assume the role of a BN variable to allow to compute probabilities on the model (c).

2.3. Learning Probabilistic Models from Ontologies

Learning a BN is an NP-hard problem whose difficulty drastically increases with the number of variables to consider. Looking at BNs, numerous related works have established that using constraints with heuristic algorithms effectively improves structure (De Campos et al., 2009; Suzuki, 1996) and parameters (De Campos & Ji, 2008; Niculescu et al., 2006) learning. In this work, the relational schema defines structural constraints as an ordering between the different variables. Node ordering yields indeed good results for BNs learning, be it a complete node ordering such as with the K2 algorithm (Cooper & Herskovits, 1992), or a partial oneas such as in (Parviainen & Koivisto, 2013). In our case, we compose a partial node ordering from expert's and ontological knowledge that we automatically transcribe in a relational schema.

Using ontological knowledge in order to learn BNs has already been proposed in several works and it is a good alternative to asking an expert inputs that can be time-consuming and prone to mistakes (Druzdzel & Gaag, 2000). Most of the existing works are based on similar methods, where the ontology

brings knowledge in order to guide the structure building. Different works combine object oriented BNs and ontologies (Ben Ishak et al., 2011; Truong et al., 2005). All these methods, however, do not include expert's knowledge, and thus do not offer a lot of flexibility. Moreover, some require the use of specific ontology's extensions. These extensions can be used to integrate probabilistic reasoning in ontologies (such as BayesOWL (Ding et al., 2006; Zhang et al., 2009) or HyProb-Ontology (Mohammed, 2016), as in (Devitt et al., 2006)). Despite their good results, all of these methods require to adapt the ontology we wish to study to these extensions, by adding new specific concepts. On the contrary, in this article, we want to focus on an existing ontology, without having to extend or modify it.

Many works are tuned to a specific ontology, but, often, without possibilities of transfer towards another domain. For instance, Bucci et al. (2011) uses predefined templates to support medical diagnosis, which cannot be extended to other medical applications; (Zheng et al., 2007; Helsper & Gaag, 2002) require to construct a specific ontology to guide the construction of the BN's structure. Our approach, despite the fact that POND is tied to a single ontology, is flexible enough to represent many different expert's cases, thanks to the genericity of the PO^2 concepts (components, steps, observations) which allow more applications than with a classical domain ontology.

In a lot of approaches at the state of the art, the BN's structure (i.e. the relations between the variables in the graph) is not learned from the data but derived from the ontology. This raises some issues as most of the ontologies do not transcribe direct causal dependencies, nor do the presence of a property indicates a correlation between the attributes it joins. For instance, (Fenz, 2012) considers that the object properties directly serve as probabilistic dependencies if they are selected beforehand by an expert. (Ben Messaoud et al., 2011) assumes that the ontology's properties are already causal in order to build an Object Oriented BN. Differently from these approaches, in our work, the pre-existence of properties can help influence the final BN, but it can never replace the statistical learning. More generally, our approach relies on statistical learning and we cannot force a relation's existence: if the data we have do not allow a relation, then we cannot draw it.

2.4. Causal Discovery from Data

One of our main motivations is to reason about the transformation process's data. This sometimes requires to deal with causality. As we have seen in Section 2.2, BNs and PRMs coupled with EG can be used for discovering

causality. However, the information represented has to be evaluated: since correlation is not causation, a data set in which we want to discover causal knowledge must answer some quality criteria. (Spirtes et al., 2000) defines the **causal sufficiency**, a set of criteria that guarantees there cannot be external factors not taken into account during the learning. (Glymour et al.) defines other criteria that affect the quality of the data used to learn causal models: it is for instance sensitive to missing data, selection bias, measurement error, non stationary or heterogeneous data and deterministic cases. In the same way that a bad data processing can lead to erroneous conclusions, if not all possible events are present in the learning set, or if their proportion is altered and does not represent reality, then it is impossible to draw good causal discoveries.

Once the data's quality has been assured, some algorithms propose to discover causality by working with independence tests between the variables (Spirtes et al., 2000; Verny et al., 2017). However, they do not allow to introduce external constraints during the learning. Other works have also proposed the use of EGs to learn causal models: Hauser & Bühlmann (2014) proposes two optimal strategies for suggesting interventions in order to learn causal models with score-based methods and the EG; Eberhardt (2008); Shanmugam et al. (2015); Castelletti & Consonni (2020) use an EG to build a causal BN while maintaining a limited number of intervention recommendations. These approaches do not require any external knowledge about the domain. In our case however, the data is structured according to an ontology and it is not sorted in a way such that a BN can be learned directly. The method we present is a continuation of Munch et al. (2019b), where expert's and ontological knowledge are used to create causal constraints during the learning to make causal discoveries. Part of this work is detailed in Section 3.3.2.

2.5. Original contributions

The original contribution of this work is to propose a complete workflow for the knowledge representation and storage of transformation processes in order to exploit this knowledge and reason about it. The workflow is detailed in the next section and includes: (1) a new version of PO^2, an ontology based on the BFO foundational ontology and SSN/SOSA W3C standards, able to represent in a very detailed way experimental transformation processes, and associated to software tools able to enrich domain ontology and annotate and query data using the ontology; (2) a mapping mechanism between

PO^2 and PRMs allowing to automatically extract comparable information from heterogeneous experimental transformation processes data; (3) an expert and ontological knowledge integration part using a PRM's relational schema building step; (4) a reasoning step allowing to analyze correlations between variables of interest, especially interesting to perform (sometimes causal) knowledge discovery. An application of this workflow, dedicated to biomass transformation processes to produce biocomposite packaging material, is presented in Section 4.

3. The POND Workflow

The functionalities of POND, the workflow proposed in this article, have been defined in the framework of several interdisciplinary projects involving computer scientists, data scientists and biomass processing experts for food and bio-based material production. In this section, we will present them from a generic point of view, not attached to a specific domain. However, it is important to note that this plurality of projects have brought us to design these functionalities with a meta-analysis scope in mind: in order to embrace this diversity, we need to be able to define and reason with data from different projects. While the PO^2 ontology allows us to define expert knowledge by unifying it under common semantic terms, reasoning about this heterogeneity requires to define specific questions that we aim to answer. We denote them as **Expert Queries (EQs)**, and separate them into two subsets:

- **Competency Questions (CQs).** In ontology engineering, CQs are natural-language questions that outline the scope of knowledge represented by an ontology and the applications exploiting it (Grüninger & Fox, 1995). CQs represent functional requirements in the sense that the ontology and the developed ontology-based information system should be able to answer them. Typical CQs adressed by PO^2 are:

  CQ_1 Which steps compose a given transformation process?

  CQ_2 Which attribute values are associated with each step?

  CQ_3 What are the attribute values associated with an input (or output) for a given step of a given transformation process?

  CQ_4 What are the changes for an attribute value of an input during a given step?

- **Knowledge Questions (KQs).** Similarly to CQs, these questions also query the knowledge modeled by the ontology and stored into the ontology-based system, but require a more in-depth analysis of the relations between the variables to deal with the uncertainty. KQs can be expressed in two different ways:

  KQ$_1$ Does a given attribute has a (causal) relation with another attribute?

  KQ$_2$ How a change in a given attribute's value (causally) influences the values' distribution of another attribute?

  The difference between CQs and KQs is that KQs do not rely on a descriptive aspect, but on the contrary require a two-times analysis, by (1) building a database representing the attributes of the question as variables and (2) learning a probabilistic model from this database to answer the question. Since they depend on this learned probabilistic model (and not directly on the ontology), KQs' definition is less domain-oriented than CQs and only rely on the concept of attributes.

More generally, CQs rely on specific concepts from the PO^2 ontology, such as the classes or properties, to be expressed; while KQs require on BNs' and PRMs' concepts such as variables that need to be define beforehand. In order to answer these, POND must implement different functionalities:

F1 The workflow should provide a representation model allowing to express both expert's and ontological knowledge useful for meta-analysis and a tool to structure and store data using such a model.

F2 In a collection of experimental data acquired during different projects, the workflow should provide a way to extract from the knowledge base, in a semi-automatic way, attributes of interest for meta-analysis purposes.

F3 The workflow should be able to compute a model able to reason with variables of interest.

If **F1** and **F2** are required for both kind of EQs, **F3** is specific to KQs. The three are provided by the following steps of the POND system, presented in Figure 2:

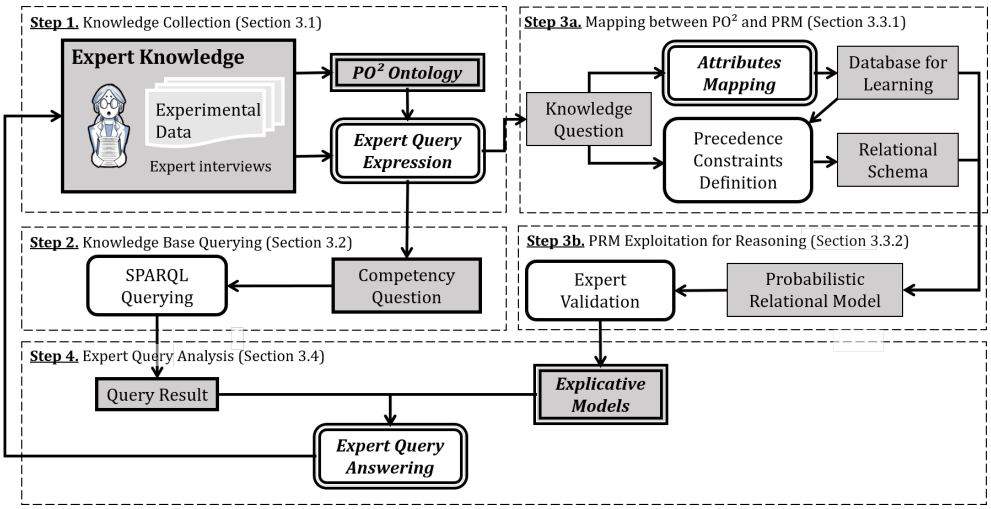

Figure 2: Workflow global overview. White boxes indicate actions that require the expert's intervention, while grey boxes indicate a concrete object automatically built from expert's inputs. Double lines show our contribution in the state of the art.

1. **Knowledge Collection** (Section 3.1). In this step, expert's knowledge is collected under the form of experimental data or expert's interviews, and structured using an ontology. PO^2 is used to annotate experimental data and to store it in a RDF database. Knowledge collection also comprises the definition of an EQs set. Depending on its type, it will either be processed in Step 2 (**Knowledge Base Querying**) for CQs or in Step 3 (PRM Learning) for KQs.

2. **Knowledge Base Querying** (Section 3.2). This step is dedicated to express CQs as **SPARQL queries** executed against the RDF database. A specific Web application, SPO^2Q, has been designed in order to assist users to query the PO^2 RDF database.

3a. **Mapping between PO^2 and PRM** (Section 3.3.1). Answering KQs requires the learning of a PRM; however, in order to integrate the expert's knowledge expressed during the **Knowledge Collection**, a mapping is first needed before interrogating the PRM. It is used to automatically translate expert's knowledge into constraints to guide the learning and it is expressed under two forms: first a mapping of the attributes, then the expression of the precedence constraints.

3b. **PRM Exploitation for Reasoning** (Section 3.3.2). Directly following the **Mapping**, this sub-step consists of the PRM's learning and its validation by the expert, who can accept or reject the result using tools to criticize the model. If the model is rejected, the expert is invited to reconsider the knowledge integration done during the **Mapping** (step 3a), and a new iteration begins. If, despite those iterations, the expert cannot validate the model, it means that the expert knowledge defined in the **Knowledge Collection** cannot be used to answer the KQ. In this case, the identified problems (such as a lack of knowledge) are given to the expert as guideline to improve the learning. On the contrary, if the model is validated, we continue to the final step, the **Expert Query Analysis**.

4. **Expert Query Analysis** (Section 3.4). This step receives the results of the SPARQL query formulated in the **Knowledge Base Querying** step and the explicative models validated in the **PRM Exploitation for Reasoning** step. These results are then analyzed to answer the EQ: for instance, if no answer has been found, we can find out whether this is due to a lack of information or a problem within one of the involved step.

Figure 2 presents these different steps and the different possible sequences. To be noted, the passage from Step 3b to either 3a or 4 depends on whether the expert has rejected or validated the learned PRM.

3.1. Knowledge Collection

This first step of the pipeline is dedicated to collect and to model the collected experimental data using an ontology and to store it into a RDF database. In this section, we will successively introduce the PO^2 ontology, and two tools developed respectively for integrating data in PO^2, PO^2 Manager, and querying data, SPO^2Q.

3.1.1. PO^2 Ontology

The PO^2 Ontology (Process and Observation Ontology) is a core model which allows us to represent a generic transformation process described by a set of experimental observations available for the inputs and outputs of each step of the production process. PO^2 contains 67 concepts and 79 relations. As presented in Section 2.1, the current version of the PO^2 model is the

upper layer for 3 domains ontologies structuring data in 3 different domains concerning food processing. The main difference between the version 1.5 of the PO^2 model published in (Ibanescu et al., 2016) is that the current version is aligned with existing standards, allowing thus interoperability:

- SSN/SOSA Janowicz et al. (2018); SOS, a lightweight ontology for Sensors, Observations, Samples and Actuators, is the result of a joint working group of the Open Geospatial Consortium (OGC) and the World Wide Web Consortium (W3C) on *Spatial Data on the Web*.

- Time Ontology TIM, a candidate recommendation of the W3C, is an ontology of temporal concepts, for describing the temporal properties of resources in the world or described in Web pages. The ontology provides a vocabulary for expressing facts about topological (ordering) relations among instants and intervals, together with information about durations and about temporal position including date-time information.

- The Basic Formal Ontology (BFO) BFO is a small, upper level ontology that is designed for use in supporting information retrieval, analysis and integration in scientific and other domains.

Figure 3 gives an excerpt of PO^2 and its relation with SSN/SOSA and the Time Ontology. The main classes of PO^2 and their relationships are:

- The `PO2:Transformation Process` class represents the sequence of steps and it is a subclass of the `sosa:Actuation` class.

- The `PO2:Step` class represents a unit operation that transforms inputs in outputs; it is a subclass of the `sosa:Actuation` class which is characterized by a temporal duration, it is performed by a device, `sosa:Actuator` (labeled by `PO2:Material`), using a given method, `sosa:Procedure` (labeled by `PO2:Method`).

- The `PO2:Observation` class is a subclass of the `sosa:Observation` class which is characterized by a temporal duration, it is performed by a sensor, `sosa:Sensor`, using a given method, `sosa:Procedure`, in order to calculate a value of an observable property, `sosa:FeatureOfInterest`; the value of the observation is a `sosa:Result` and concerns a property, `ssn:Property`, labeled by `PO2:Attribute`.

- The `PO2:Component` class represents entities, inputs and outputs of a step; it is a subclass of the `sosa:Feature of interest` class.

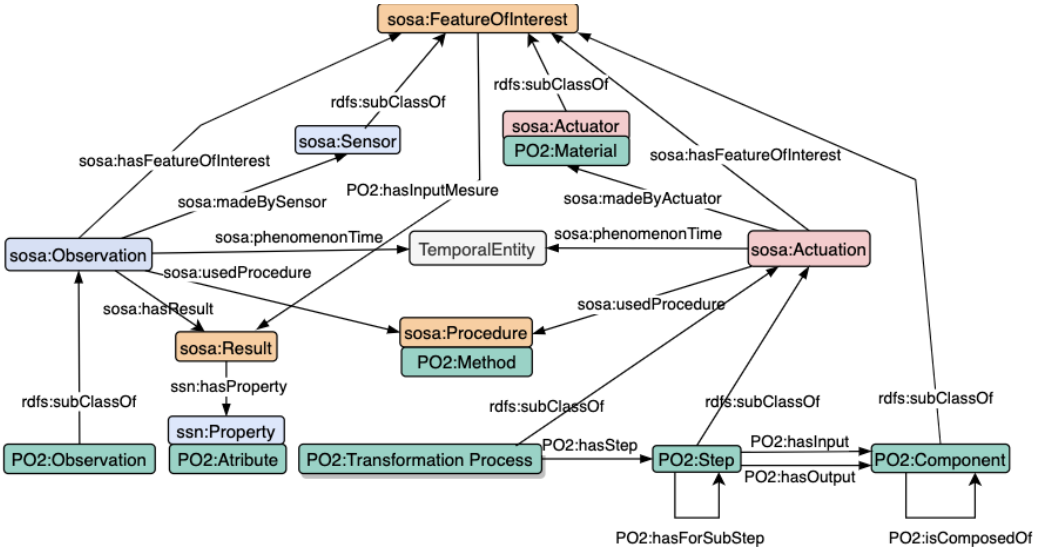

Figure 3: PO^2 ontology.

When designing a new domain ontology based on the PO^2 core model, the seven classes prefixed with `PO2:` in Figure 3 should be specialized with concepts from the new domain. This task may be done in a convenient way with the specific tool, PO^2 Manager, presented in Section 3.1.2.

PO^2 ontology version 2.0, implemented in OWL 2 OWL, is published on the AgroPortal ontology library PO2 (b), and is Creative Commons Attribution International (CC BY 4.0) CCB. The last version of PO^2 ontology, version 2.2.1, is available on its home page (PO2, a).

3.1.2. PO^2 Manager

PO^{2}Manager (Dervaux et al., 2018; Buche et al., 2020), a standalone application developed in Java, is a tool designed to assist domain experts on two tasks: (i) extending the PO^2 core concepts with concepts associated with the application domain; (ii) describing experimental processes using the concepts of the application domain.

PO^{2}Manager assists users in specializing the core concepts of the PO^2 ontology in order to allow annotations of processes in a given application domain. For example, in Figure 4, the *Softwood* concept is defined as a special-

ization of the PO^2 core concept *component*. As PO^{2}Manager is able to scan ontology portals as AgroPortal (Jonquet et al., 2018), it is able to automatically align the concept belonging to the PO^2 domain ontology with concepts already defined in available ontologies. In Figure 4, the *Softwood* concept is associated with four concepts defined in NALT (NAL), Agrovoc (Agr), CABI and GACS (Baker et al., 2019) using the `skos:exactMatch` property.

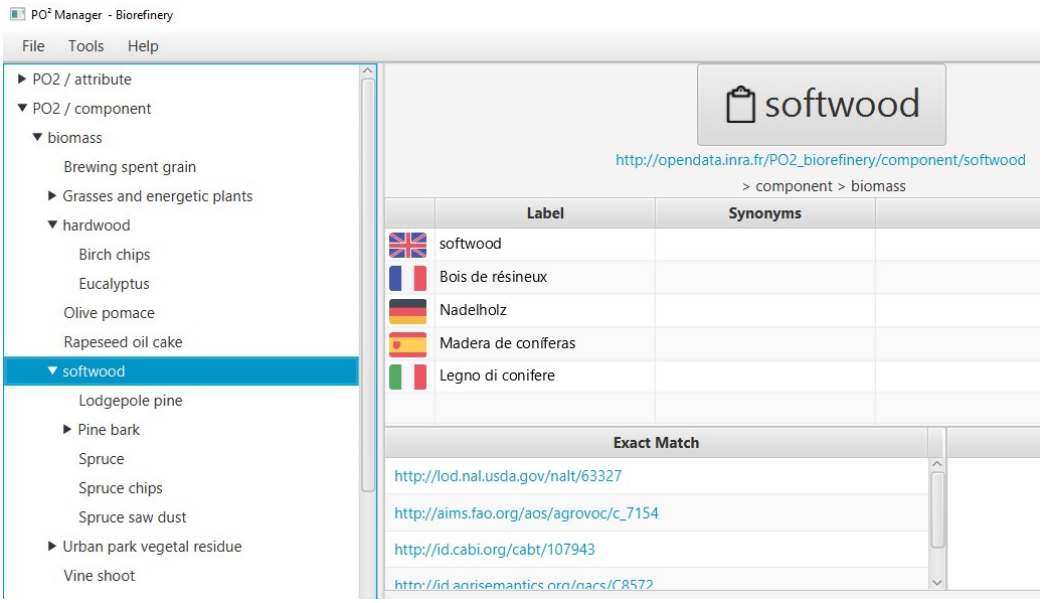

Figure 4: Domain ontology concept edition with PO^{2}Manager

A graphical user interface (GUI) has been designed in PO^{2}Manager to allow the manual entering of processes. The left part of Figure 5 shows the editor which allows to create processes composed of unit operations and input and output components. For example, in Figure 5, the process *Itinerary 20* is composed of three sequential unit operations (namely a *centrifugal milling* followed by an *extrusion* and a *film calendering*). *Pinepark1* is an input component of the centrifugal milling and *CE* is a powder which is the output component of the milling operation. The right part of Figure 5 shows the graphical representation of the process. Moreover, observations may be associated with unit operations and characterised by measures. In Figure 6, mechanical characteristics (e.g. stress at break, Young modulus, etc.) associated with the output component named *2% Pine bark fiber/PHBV film* of the film calendering operation are shown.

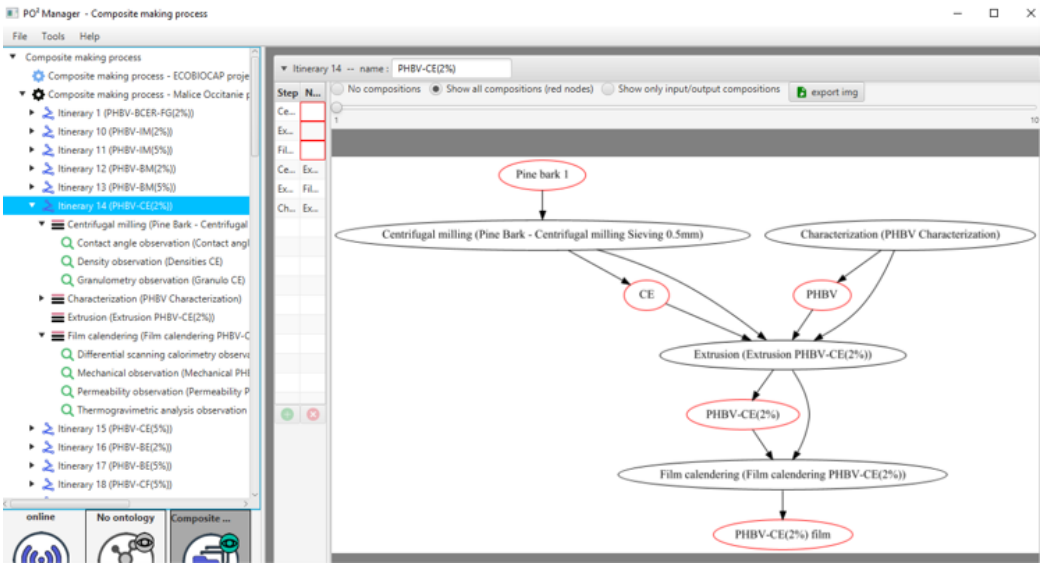

Figure 5: Process edition with PO²Manager

3.2. Knowledge Base Querying

SPO²Q (Buche et al., 2020) is a Web application designed to assist users to query the PO² RDF database through its SPARQL endpoint. A set of SPARQL queries are pre-defined for users which are not familiar with SPARQL. Those pre-defined queries can be specialized with concepts of the application domain. In an advanced usage of SPO²Q, complex SPARQL queries may be defined. An example of use is given in Section 4.2, where Figure 8 and 9 illustrate CQ_2 (presented in Section 3) and its answer, within a specific domain of application.

3.3. PRM Learning

As we have seen, KQs address the data represented in PO² on a higher reasoning level than CQs that considers relations between variables to deal with uncertainty. They need probabilistic models to be answered. In this section, we will present the two steps needed to learn those: a first about the automatic mapping between PO² and the PRMs, and a second about the PRM exploitation for reasoning.

3.3.1. Mapping between PO² and PRM

Since KQs don't directly rely on the ontology semantics, in addition to the knowledge already encapsulated in the knowledge base, each KQ requires

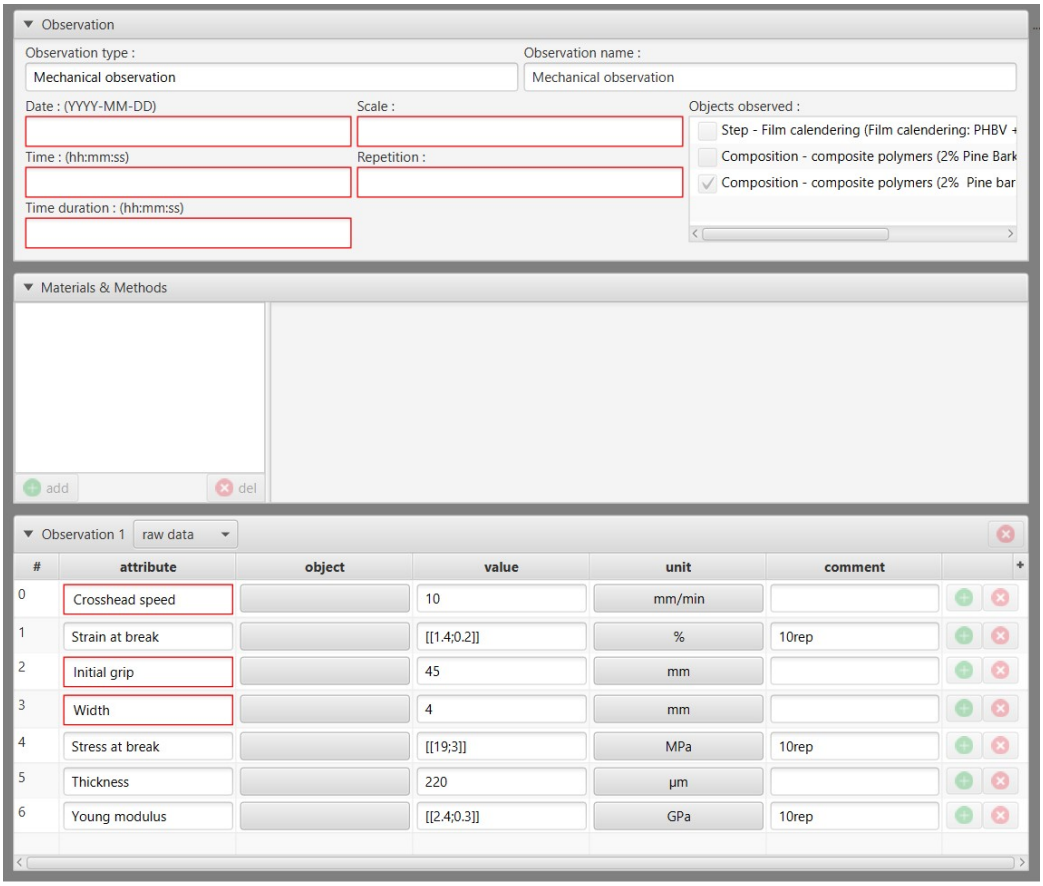

Figure 6: Edition of characteristics associated with a sample in PO^{2}Manager

extra inputs for being answered. Due to their problem-dependent intrinsic nature, these inputs cannot be formalized in knowledge bases. They however are required to narrow the learned model with the problem of interest. To do so, the expert is invited to provide, for each KQ, inputs of two different forms:

- **Attributes Mapping.** Building the database to learn the model requires to associate to each variable the corresponding attributes in the knowledge base. If this problem is sometimes straight-forward (i.e. it is easy to find a corresponding attribute), it requires sometimes more information. For instance, one can ask to reason about a "quantity of flour"; but depending of the recipe, flour may be incorporated in

multiple times. In this specific case, there is thus a need to define "the total quantity of flour added", which might not exist in the knowledge database. Thanks to this first step, the expert can specify how to define each variable and which attributes can be associated to them in order to automatically build the learning database.

- **Precedence Constraints Definition.** When learning the relations between different variables, the algorithm pre-supposes that there are no constraints between them: given A and B two attributes, both relations A $\rightarrow$ B and B $\rightarrow$ A can be learned. On the contrary, by defining precedence constraints, the expert can force an orientation. This integration of knowledge guides the learning towards a result closer to the reality. Moreover, if these constraints are considered as causal, then it creates a context favorable to causal discovery (Munch et al., 2019a).

If precedence constraints definition can be easily done thanks to the PRM's structure and definition of the relational schema (whose relational slots translate precedence constraints among variables), there is, to our knowledge, no way to automatically map concepts from an ontology (in our case, the attributes from PO^2) to variables of a probabilistic model. In the rest of this section, we will cover this specific problem by introducing our solution with a small toy example. Be Process_1 a transformation process realized through two different itineraries (see Fig. 7):

- Itinerary_1. A succession of two steps, Step_1 and Step_3, respectively associated to two outputs: Component_1 and Component_3.

- Itinerary_2. A succession of two steps, Step_2 and Step_3, respectively associated to two outputs: Component_2 and Component_3.

These two itineraries both have Step_3 and Component_3: since they bear the same name, we can infer that they semantically represent the same concept. This is not the case, for instance, for Step_1 and Step_2: even if they both precede Step_3, their names differ, and thus they are considered as two distinct concepts. This raises the question of the comparability of the measures: is the value of temperature of Component_1, measured during Step_1, of the same nature than the one of Component_2, measured during Step_2? This kind of knowledge cannot be natively added in the ontology, hence the need for the expert's definition of variables. For the rest of this example,

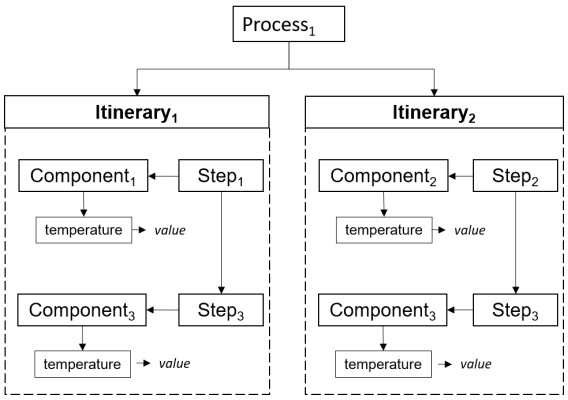

Figure 7: Example of a small transformation process of two different itineraries. This representation uses a simplified version of the ontology's semantic to facilitate the reading.

we will consider that the temperatures of Component_1 and Component_2 are comparable.

To help the mapping, we ask the expert to indicate, for each variable of the KQ, the corresponding attribute(s) in the knowledge base and its (their) location(s), with the help of the ontology to select the interesting values. Using the ontological concepts of Process, Step, Observation and Attribute, inherent to the PO^2 semantics has two advantages:

- The expert can easily pinpoint the location of the interesting values for building the variable while relying on known and well-defined vocabulary
- SPARQL queries can be automatically generated using these concepts to build the database necessary for the learning.

Using the transformation process described in Figure 1, we define the variables described in Table 1. Each variable is differentiated from the other thanks to the variable's label to which it is attributed. This way, multiple attributes can correspond to a same variable. In this example, we have defined two variables corresponding to two temperatures: one measured during the first step (either Step_1 or Step_2), and another measured during the second step (Step_3).

3.3.2. PRM Exploitation for Reasoning

Once the attributes mapped to variables and precedence constraints defined, a PRM can be learned. Yet, while we aim to learn a model representing

Process	Step	Observation	Attribute	Variable's Label
Process_1	Step_1	Component_1	Temperature	Temperature_1
Process_1	Step_2	Component_2	Temperature	Temperature_1
Process_1	Step_3	Component_3	Temperature	Temperature_2

Table 1: Table used for mapping PO^2's attributes to probabilistic model's variables.

the reality, we have however no way to directly evaluating the performances of the learned result. That is why, before using the PRM to answer the KQ, we submit it to an expert's validation phase. The main idea is to submit the learned relations to the expert and verify whether (1) they corroborate them and (2) causality can be deduced from them. It is important to note that even if the KQ does not require causal discovery to be answered, this part is important for the model validation as it gives the expert tools to criticize the learned model. If a learned relation is causally contradicting expert knowledge, then it indicates that the model cannot be exploited as such. Moreover causal reasoning can only be done if we consider the causal criteria (such as the causal sufficiency) defined in Sec. 2.4. In our case, the resulting model can be seen as the intersection between all the models constrained by the dataset (expressed by the EG) and all the models constrained by the expert's causal knowledge (expressed by the RS). When looking at both the EG and the RS, we can deduce for each relation whether its causality is validated or not, following Munch et al. (2019b):

- If the relation is oriented in the EG, then the causality of this arc is validated by the EG.
- If the relation has been influenced by a precedence constraint, then its orientation is due to an expert intervention. Thus, its causality is validated by the expert.
- If the relation is neither influenced by a precedence constraint nor oriented in the EG, then it is impossible to deduce its causality.

In order to answer the KQ, the expert has to verify and validate each relation. In the case where one or more relations are not validated, the expert has to reject the learned PRM and return to the **Mapping** step. There, they can modify the pool of variables, or specify new precedence constraints. This back and forth continues until either the PRM is finally validated, or the database is deemed unfit for answering the KQ.

3.4. Expert Query Analysis

If this step is common for all EQ, it is important to note that due to the method's differences, answering a CQ is not the same as answering a KQ.

3.4.1. Answering Competency Questions

In the case of CQ, the SPARQL query has provided a set of results. Depending on it, the CQ can then have no answer (i.e., the set of results is empty), a unique one (i.e., the set's length is 1), or multiple possible (i.e., the set's length is strictly superior to 1). This answer is then reported to the expert.

3.4.2. Answering Knowledge Questions

In the case of KQ, the answer can only be computed once the learned PRM has been validated by the expert. Then, depending of the KQ, the complexity of the answer might vary:

- **Answering KQ_1.** In order to check whether an attribute has a relation with another, it is sufficient to check if a path exists between the two attributes in the learned model. If not, then they are independent.
- **Answering KQ_2.** In order to check how a change in an attribute's values influences the values' distribution of another attribute, we can use the conditional probability table to see whether a change of the first's values modifies the second.

Causal knowledge deduced in the previous section can also be used to answer more specific KQ about causal relations. For instance, if the causality of the relation $\text{Attribute}_1 \rightarrow \text{Attribute}_2$ has been validated, then it means that changing the value of Attribute_1 will have an influence over Attribute_2. This case will be illustrated in the following section.

4. Results

In this section, we present a real world application on the processing of biocomposites for food packaging materials. Nowadays, the amount of plastic used each year raises a substantial number of questions about their harmful impact on the environment, the eco-systems and human health. In order to find substitutes, poly(3-hydroxybutyrate-co-3-hydroxyvalerate), called PHBV, is a promising bacterial bio-polymer that is biodegradable in soil and

ocean and that can be synthesized from all kinds of carbon residues. The potential for a quite larger PHBV market will be ensured provided that their mechanical and thermal properties can be further improved, and their cost and environmental impacts reduced. The development of biocomposites via the incorporation of lignocellulosic fillers (LFs) obtained by dry fractionation of unrecovered organic residues (agricultural, urban, forestry and from agro-food industries) was proved to reduce the overall cost and environmental impact of PHBV-based composite materials (David et al., 2020a), while modulating their functional properties (David et al., 2020b; Berthet et al., 2015b). However, the introduction of lignocellulosic fibers has a negative impact over the biocomposite's brittleness and process-ability (Berthet et al., 2015a). It was shown that decreasing the lignocellulosic filler size down to several μm allowed to optimize both the cost (maximal highest possible filler content) and the functional properties of biocomposites. Therefore, the goal of biocomposites' processing is to find the right compromise between the maximum acceptable filler content, the filler size and the resulting properties.

4.1. Domain Description

We collected data from three projects focused on developing PHBV-based biocomposites using lignocellulosic fillers (LFs) stemming from organic waste streams, e.g. crop residues (*Chercheur d'avenir region Languedoc-Roussillon MALICE*, *H2020 NoAW*) and urban waste (*H2020 Resurbis*). These interdisciplinary projects involving computer scientists, data scientists and biomass processing experts for food and bio-based material production, provided a unique set of experimental data concerning the production of new bio-materials, with 70 different formulations. Available data concern the technical process descriptions, including the description of each step, its inputs and outputs and the description of different possible itineraries. Some specific attributes (illustrated in the next sections) are observed/measured during some specific steps or itineraries. The final database presents 142,688 triples.

In order to elicit the best formulation to produce biocomposite, the experts want to inquire the impact of the incorporated LFs to the PHBV's attributes and, more generally, find the optimal combination that improves or better preserves the final properties of PHBV-based materials while decreasing their overall cost and environmental impact. The above research

questions may be expressed either as CQs or KQs. In the rest of this article, we will present how POND can help to answer them.

4.2. Competency Question Answering

We illustrate this part by answering a specific CQ derived from CQ_2 ("Which attribute values are associated with each step?", introduced in Section 3), which can be executed with the form presented in Figure 8. This corresponds to the interrogation: "Which values are associated to the attribute "Rotation speed" in the Step *Vine shoot - Impact milling 100 UPZ sieving size 0.3 mm?*". Associated results are presented in Figure 9: in all itineraries in which the impact milling step has been involved, rotation speed was always fixed to 18,000 min-1. It can be modified by the user to express more elaborated queries than those which can be generated from the form using SPOOQ in (SPARQL) expert mode. More generally, POND allows to answer any CQ as presented in the introduction of Section 3.

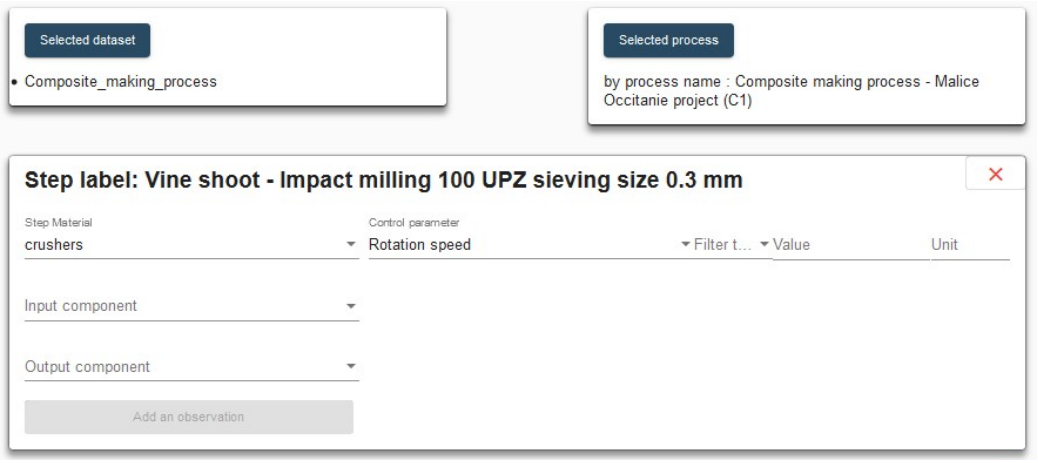

Figure 8: Executing CQ2 with SPOOQ

process_1	sampleNameLabel_1	process_sample_code_1	itineraryLabel	step_1	stepMaterial_1	materialParameter_1	valueOrigin_matParam_1	unitOrigin_matParam_1
Composite making process - Malice Occitanie project	Composite	C1	Itinerary 4 - Itinerary 4 composite with 2% vine shoot fibers	Vine shoot - Impact milling 100 UPZ sieving size 0.3 mm	crushers	Rotation speed	18000	min-1
Composite making process - Malice Occitanie project	Composite	C1	Itinerary 5 - Itinerary 5 composite with 5% vine shoot fibers	Vine shoot - Impact milling 100 UPZ sieving size 0.3 mm	crushers	Rotation speed	18000	min-1
Composite making process - Malice	Composite	C1	Itinerary 2 - Itinerary 2 composite with 5% vine	Vine shoot - Impact milling 100 UPZ	crushers	Rotation speed	18000	min-1

Figure 9: Results associated with CQ2 in SPOOQ

4.3. Knowledge Question Answering

In this part, we will answer the following KQ:

KQ$_{bio}$ Which parameters explain the thermal degradation temperatures?

In order to do the **Attributes Mapping**, we select the following attributes from the RDF database and map them into variables[3]:

- **Variables that describe the LFs.** LFs are described by three main categories of variables: biochemical composition, i.e. ***cellulose***, ***hemicellulose***, ***ash*** and ***lignin*** (in %); apparent median diameter ***d50*** (in μm); weight ***filler content*** (in wt%, that represents the proportion of incorporated LF).

- **Variables that describe the virgin PHBV and the PHBV-based biocomposites.** Those are described by four different categories of variables: tensile properties (***strain at break***, ***stress at break*** and ***young's modulus***), ***permeability*** (to water vapour), thermal properties (***crystallization*** and ***melting*** **temperatures**) and thermal degradation temperatures (***onset*** and ***peak*** temperatures). For each set of biocomposites, all the properties were normalized with respect to the properties of the corresponding PHBV matrix.

This represents in the end a database of 84 samples of biocomposites with 15 variables to describe their components and their final properties. The expert can then organize them in the relational schema presented in Figure 10. Once the PRM learned, it is instantiated as the BN presented in Figure 11. As we can see, the ***peak*** degradation temperature is explained here by the ***filler content*** and the ***ash*** content, while the ***onset*** degradation temperature is only explained by the ***filler content***. The causality of these links is enforced by the relation schema's precedence constraints, previously established by the expert: it is considered here that an attribute defining the input can have a causal impact over an attribute defining the output, but not the contrary. Therefore, we can easily answer KQ$_{bio}$ by looking at the learned model, without having to consult the EG - which would have been relevant for intra-class relations. This shows that the parameters having an

[3]For the rest of the article, all variables represented in the database are ***emphasized***.

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

Table 2: Conditional Probability Table showing the influence of the **Filler Content** over the **Peak** Temperature distribution. * shows the maximum likelihood.

influence over the thermal degradation temperatures are the ***filler content*** and the ***ash*** content.

In order to validate these results, we compare the answer with the domain's state of the art:

- ***Filler Content*** $\rightarrow$ ***Peak/Onset*** shows that when increasing the ***filler content***, the average temperature tends to decrease (Table 2), which is indeed verified in multiple state of the art works (David et al., 2020b). This verifies the premise introduced in the previous section: the goal of optimizing bio-sourced packages processes is to find the right compromise between the ***filler content*** (to decrease the cost of the PHBV) and the preservation of the mechanical and thermal properties.

- ***Ash*** $\rightarrow$ ***Peak*** temperature, on another hand, has not been verified by any work. In order to validate (or refute) it, more experiments should be realised to test this hypothesis.

In the end, the expert can use the learned model to reason on possible compromises, depending on what they prefer. Table 2 for instance presents the interaction between the ***filler content*** and the ***peak temperature***. If the goal is to augment the ***filler content*** to decrease the overall costs, while not decreasing too much the ***peak temperature***, then choosing a filler at]4;11] might be the best compromise. However, if the preservation of the ***peak temperature*** is more important, then choosing a filler at]2;4] is more interesting, as it guarantees the highest probability (0.82) of having an almost non-degraded temperature.

More generally, a same reasoning can be applied to each characteristic of the composite for assessing the best LFs and the optimal filling rate. Figure 11 shows that the LF's composition has an impact over the mechanical

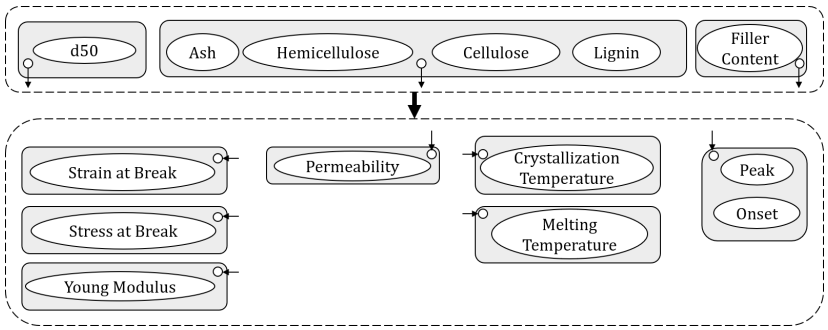

Figure 10: Relational Schema defined for answering KQ_{bio}. To increase the readability, arrows have been summarized: each of the three upper classes has a precedence constraint directed towards the lower classes. This symbolizes the influence of the input's attributes over the output's attributes.

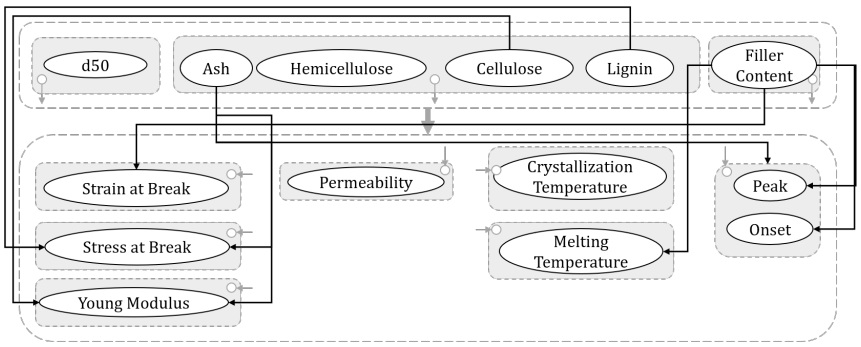

Figure 11: BN instantiated from the PRM learned using the relational schema defined in Figure 10.

properties and the ***peak temperature***; and that the ***filler content*** has a impact over the ***melting***, ***peak*** and ***onset*** temperatures, as well as the ***strain at break***.

5. Discussion

In this section, we will cover the discussion over both the overall method, and the illustration case.

5.1. POND Workflow Discussion

In Section 3 we have presented the different parts of the workflow. As introduced, the first originality of our approach stems from the combination of

different methods whose efficiency has been demonstrated independently in state-of-the-art works. The restriction over a single ontology, PO^2, although limiting on the domains that can be studied using POND, allows a better automation and optimisation of the interactions with the expert. In particular, it helps us targeting the EQs that POND aims to answer. Those are distributed between CQs (that had already been defined within the ontology domain), and KQs (that we defined especially for POND). While this global description of CQs tends to be general enough for most cases, it might be possible that future works will bring to light some new expert's interrogations, that are not covered by our current definition.

In the same manner, when answering KQs, the integration of expert's knowledge has been defined in two different ways, which is a second originality of our approach: the mapping of the relevant attributes into variables on one hand, and the integration of precedence constraints on another. The use of PO^2's vocabulary when integrating this new information allows an automatic building of the learning database through well-defined SPARQL queries. The main limit of answering a KQ, however, lies on its deep ties with both the dataset and the expert's contribution. We have presented in Section 3.3.2 the importance of expert's validation. However, the results learned can be greatly influenced either by the dataset (usually provided by the expert themselves) or the expert's causal constraints. Moreover, the choice of discretization (made during the variable selection part) can also have a great impact on the final result. As a consequence, all learned models have to be considered true under specific conditions, such as the causal sufficiency (presented in Section 2.4) and the assurance that the knowledge represented in the dataset and provided by the expert is true and representative.

The third originality of this paper is the development of a new version of PO^2, built upon ontological standards. It is accompanied by two tools developed to implement it.

5.2. Results Discussion

The application on the processing of biocomposites for food packaging covers the two kinds of EQs. It first presents different kinds of CQ, as well as the detailed answer for one of them. In a second time, it presents a specific KQ, and the walk-through from this query to the learning of a model dedicated to its answer. In order to showcase POND's results, we validate the learned model with a comparison to the state of the art. This validation has two phases:

- **Validation of the relation between the thermal degradation temperatures and the filler content.** This relation has indeed been developed in recent state of the art (David et al., 2020b), highlighting the fact that the more the filler content, the lower the temperatures of thermal degradation are. It is interesting to note however that the probability of having the lowest peak temperature is higher for a filler content ranging from 11 to 21 wt% (probability of 0.91) than for a filler content ranging from 21 and 50 wt% (probability of 0.6) (Table 2). If these results are true, this would mean that the probability of obtaining a higher temperature augments with the filler content, which is not right. This result can be discussed considering that the probability of obtaining the lowest temperature at]11;21] is quasi certain (i.e. the other probability is near zero). This would tend to indicate that some cases were not present in the learning dataset, and thus were considered impossible by the algorithm. In order to confirm (or infirm) this observation, more experimental data should be added in the dataset, especially at high filler contents. Consequently in this case, the method identifies possible knowledge gaps which is itself a relevant result for the expert even if it was not defined as an EQ.

- **Validation of the relation between the peak thermal degradation temperature and the ash content.** As presented in the results analysis, this relation does not echo any work from the state of the art. This would tend to indicate either a bias in the dataset (that would enforce a spurious correlation between the two variables), or a new to-be-discovered knowledge. In both cases, new experiments should be added to the dataset in order to verify this relation.

Considering the quantity of data needed to learn robust models, both of these conclusions are not surprising. It is however interesting to note that, despite this lack of precision, the general result can still be validated by the state of the art. It illustrates another goal for POND, which is the questioning of the dataset and the indication of potential enrichment.

6. Conclusion

In this article, we have presented POND, a workflow dedicated to the representation and the knowledge discovery for transformation processes. After presenting the main definitions for describing a process, we have introduced

the two components on which this workflow is built: PO^2, an ontology used for describing transformation processes; and the PRMs, a probabilistic model able to intuitively integrate expert's knowledge during its learning. This workflow is divided in different parts: the knowledge collection, the CQ answering (through the Knowledge Base Querying), and the KQ answering (through the mapping between PO^2 and the PRM and the exploitation of this for reasoning). These are divided between the different tools and methods presented throughout this article, and help the expert to gain a new overview of the studied domain, for instance by suggesting new experiments or ways to improve the dataset.

This new insight could, for instance, be used to link the knowledge base with other ontologies that could bring new ways to enrich the dataset. For instance, Figure 11 shows that the **Stress at Break** and the **Young Modulus** depend on the composition of the used biomass (more specifically, their composition of **Ash**, **Cellulose** and **Lignin**). Using other ontologies, such as one presenting new different biomasses, it could be interesting to look for new propositions that correspond to the values recommended by the learned models and propose new experiments to the expert. In a more general setting, it should be interesting to investigate how other ontologies could be linked to the studied domain given the analysis provided by POND. This could help establishing matches between ontologies representing different aspects of a transformation process's domain.

On another hand, POND could also be used to generate rules (such as SHACL constraints) that evaluate the data represented in the ontology. For instance, these rules can be used to define a set of constraints that can evaluate the quality of potential new data (probable, not probable, etc.) and help the expert find outliers. Since our model was learned using causal knowledge, these new constraints would be motivated by causal relations and then be easily explainable.

Moreover, as we have seen, the quality of the database has a strong influence over the final model. This could become problematic if it introduces biases in the result. For now, POND considers all sources as equivalent for the learning; however, by enriching the knowledge graph with indication of the data source and its quality, it becomes possible to weight the different inputs in order to incorporate this reliability in the learning, by favoring data whose source is more trustworthy.

Finally, we have only considered cases where expert agreed with each other (i.e. there are no point of disagreement on the studied subject). How-

ever, it is possible that during the knowledge collection part (such as for the relational schema building) POND is given multiple contradictory information. It should be interesting to develop more our framework to better handle such cases.

7. Acknowledgement

This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 688338 and the FUI23 Meatyl@b project financed by BPI France under grant DOS0058786/00.

References

(). Basic Formal Ontology (BFO). https://github.com/BFO-ontology/BFO. Accessed: 2022-01-18.

(). BioPortal - a repository of biomedical ontologies. https://bioportal.bioontology.org/. Accessed: 2022-01-18.

(). Core Ontology for Biology and Biomedicine (COB). http://www.obofoundry.org/ontology/cob.html, note = Accessed: 2022-01-18.

(). Creative Commons Attribution International 4.0 International (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/. Accessed: 2022-01-18.

(). Extensions to the Semantic Sensor Network Ontology. (https://www.w3.org/TR/vocab-ssn-ext/). Accessed: 2022-01-18.

(). Information Artifact Ontology (IAO). https://github.com/information-artifact-ontology/IAO/,. Accessed: 2022-01-18.

(). Linked Open Vocabularies (LOV). https://lov.linkeddata.es/dataset/lov. Accessed: 2022-01-18.

(). Open Geospatial Consortium (OGC). https://www.ogc.org/. Accessed: 2022-01-18.

(a). Process and Observation Ontology . http://quantum.agroparistech.fr/PO2. Accessed: 2022-01-18.

(b). Process and Observation Ontology . http://agroportal.lirmm.fr/ontologies/PO2. Accessed: 2022-01-18.

(). Relation Ontology (RO). http://www.obofoundry.org/ontology/ro.html. Accessed: 2022-01-18.

(). Semantic Sensor Network Ontology (SSN). https://www.w3.org/TR/2017/REC-vocab-ssn-20171019/. Accessed: 2022-01-18.

(). The FAO AGROVOC Thesaurus. https://www.fao.org/agrovoc/. Accessed: 2022-01-18.

(). The International Framework for Food Description. https://www.langual.org//. Accessed: 2022-01-18.

(). The Linked Open Data Cloud. https://lod-cloud.net/. Accessed: 2022-01-18.

(). The National Agricultural Library Thesaurus (NALT). https://data.nal.usda.gov/dataset/nal-agricultural-thesaurus-and-glossary. Accessed: 2022-01-18.

(). The Open Biological and Biomedical Ontology (OBO) Foundry. http://obofoundry.org/. Accessed: 2022-01-18.

(). Time Ontology in OWL. (https://www.w3.org/TR/owl-time/). Accessed: 2022-01-18.

(). Web Ontology Language (OWL). https://www.w3.org/2001/sw/wiki/OWL. Accessed: 2022-01-18.

(). World Wide Web Consortium (W3C). https://www.w3.org/. Accessed: 2022-01-18.

Baader, F., Horrocks, I., Lutz, C., & Sattler, U. (2017). *An Introduction to Description Logic*. Cambridge University Press.

Baker, T., Whitehead, B., Musker, R., & Keizer, J. (2019). Global agricultural concept space: lightweight semantics for pragmatic interoperability. In *npj Science of Food*. volume 3.

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., Jensen, M., Lin, Y., Lister, A. L., Lord, P., Malone, J., Manduchi, E., McGee, M., Morrison, N., Overton, J. A., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Schober, D., Smith, B., Soldatova, L. N., Stoeckert, C. J., Jr., Taylor, C. F., Torniai, C., Turner, J. A., Vita, R., Whetzel, P. L., & Zheng, J. (2016). The ontology for biomedical investigations. *PLOS ONE*, *11*, 1–19. doi:10.1371/journal.pone.0154556.

Ben Ishak, M., Leray, P., & Ben Amor, N. (2011). Ontology-based generation of object oriented bayesian networks. *CEUR Workshop Proceedings*, *818*, 9–17.

Ben Messaoud, M., Leray, P., & Ben Amor, N. (2011). Semcado: A serendipitous strategy for learning causal bayesian networks using ontologies. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, (pp. 182–193).

Berthet, M.-A., Angellier-Coussy, H., Chea, V., Guillard, V., Gastaldi, E., & Gontard, N. (2015a). Sustainable food packaging: Valorising wheat straw fibres for tuning phbv-based composites properties. *Composites Part A: Applied Science and Manufacturing*, *72*, 139–147.

Berthet, M.-A., Angellier-Coussy, H., Machado, D., Hilliou, L., Staebler, A., Vicente, A., & Gontard, N. (2015b). Exploring the potentialities of using lignocellulosic fibres derived from three food by-products as constituents of biocomposites for food packaging. *Industrial Crops and Products*, *69*, 110–122.

Bizer, C. (2013). Interlinking scientific data on a global scale. *Data Sci. J.*, *12*, GRDI6–GRDI12.

Bucci, G., Sandrucci, V., & Vicario, E. (2011). Ontologies and bayesian networks in medical diagnosis. *Proceedings of the Annual Hawaii International Conference on System Sciences*, (pp. 1–8).

Buche, P., Cufi, J., Dervaux, S., Dibie, J., Ibanescu, L., Oudot, A., & Weber, M. (2020). Food transformation process description using PO^2 and FoodOn. *Integrated Food Ontology Workshop (IFOW) ICBO*, .

Buche, P., Dibie-Barthélemy, J., Ibanescu, L., & Soler, L. (2013). Fuzzy web data tables integration guided by an ontological and terminological resource. *IEEE Trans. Knowl. Data Eng.*, *25*, 805–819.

Castelletti, F., & Consonni, G. (2020). Discovering causal structures in bayesian gaussian directed acyclic graph models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, .

Chickering, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, *3*, 507–554. URL: `https://doi.org/10.1162/153244303321897717`. doi:`10.1162/153244303321897717`.

Cooper, G. F., & Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, *9*, 309–347.

David, G., Croxatto Vega, G., Sohn, J., Nilsson, A. E., Hélias, A., Gontard, N., & Angellier-Coussy, H. (2020a). Using life cycle assessment to quantify the environmental benefit of upcycling vine shoots as fillers in biocomposite packaging materials. *International Journal of Life Cycle Assessment*, .

David, G., Vannini, M., Sisti, L., Marchese, P., Celli, A., Gontard, N., & Angellier-Coussy, H. (2020b). Eco-Conversion of Two Winery Lignocellulosic Wastes into Fillers for Biocomposites: Vine Shoots and Wine Pomaces. *Polymers*, *12*, 1530.

De Campos, C., Zeng, Z., & Ji, Q. (2009). Structure learning of bayesian networks using constraints. *Proceedings of the 26th Annual International Conference on Machine Learning, ICML'09*, (pp. 113–120).

De Campos, C. P., & Ji, Q. (2008). Improving bayesian network parameter learning using constraints. *19th International Conference on Pattern Recognition*, (pp. 1–4).

Dervaux, S., Dibie, J., & Ibanescu, L. (2018). Po^2 vocabularymanager - a collaborative tool to assist users in building a po^2 domain ontology linked with existing resources. *SemFAEN 2018 : Semantic for Food, Agriculture, Environment and Nutrition*, .

Devitt, A., Danev, B., & Matusikova, K. (2006). Constructing bayesian networks automatically using ontologies. *Applied Ontology*, *1*.

Ding, Z., Peng, Y., & Pan, R. (2006). Bayesowl: Uncertainty modeling in semantic web ontologies. *Soft Computing in Ontologies and Semantic Web*, (pp. 3–29).

Doan, A., Halevy, A. Y., & Ives, Z. G. (2012). *Principles of Data Integration*. Morgan Kaufmann.

Dooley, D., Weber, M., Ibanescu, L., Lange, M., Chan, L., Soldatova, L., McGinty, H. K., Yang, C., & Hsiao, W. (2021). Food process ontology requirements. *IFOW 2021 Integrated Food Ontology Workshop, September 15-18, Bolzano Italy*, .

Dooley, D. M., Griffiths, E. J., Gosal, G., Buttigieg, P. L., Hoehndorf, R., Lange, M., Schriml, L. M., Brinkman, F. S. L., & Hsiao, W. W. L. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. In *npj Science of Food* (p. 23). volume 2.

Druzdzel, M., & Gaag, L. C. (2000). Building probabilistic networks: Where do the numbers come from? *IEEE Transactions on Knowledge and Data Engineering*, *12*, 481–486.

Eberhardt, F. (2008). Almost optimal intervention sets for causal discovery. *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, (pp. 161—-168).

Fenz, S. (2012). An ontology-based approach for constructing bayesian networks. *Data & Knowledge Engineering*, *73*, 73–88.

Friedman, N., Getoor, L., Koller, D., & Pfeffer, A. (1999). Learning probabilistic relational models. *International Joint Conference on Artificial Intelligence*, (p. 1300–1309).

Gangemi, A., & Presutti, V. (2009). Ontology design patterns. In *Handbook on Ontologies* (pp. 221–243).

Getoor, L., & Taskar, B. (2007). *Introduction to statistical relational learning*. The MIT Press.

Glymour, C., Zhang, K., & Spirtes, P. (). Review of causal discovery methods based on graphical models, .

Gruber, T. (2016). Ontology. In L. Liu, & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 1–3). New York, NY: Springer New York.

Grüninger, M., & Fox, M. S. (1995). The role of competency questions in enterprise engineering. In A. Rolstadås (Ed.), *Benchmarking — Theory and Practice* (pp. 22–31). Boston, MA: Springer US.

Guarino, N., Oberle, D., & Staab, S. (2009). What is an ontology? In S. Staab, & R. Studer (Eds.), *Handbook on Ontologies* International Handbooks on Information Systems (pp. 1–17). Springer Berlin Heidelberg.

Hauser, A., & Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, *55*, 926—-939.

Helsper, E., & Gaag, L. C. (2002). Building bayesian networks through ontologies., . (pp. 680–684).

Ibanescu, L., Dibie, J., Dervaux, S., Guichard, E., & Raad, J. (2016). PO^2 - A Process and Observation Ontology in Food Science. Application to Dairy Gels. In E. Garoufallou, I. S. Coll, A. Stellato, & J. Greenberg (Eds.), *Metadata and Semantics Research - 10th International Conference, MTSR 2016, Göttingen, Germany, November 22-25, 2016, Proceedings* (pp. 155–165). volume 672 of *Communications in Computer and Information Science*. doi:10.1007/978-3-319-49157-8.

Ireland, J. D., & Møller, A. (2000). Review of International Food Classification and Description. *Journal of Food Composition and Analysis*, *13*, 529–538.

Ireland, J. D., & Møller, A. (2010). Langual food description: a learning process. *Eur J Clin Nutr*, *64*, 44–48.

Ireland, J. D., & Møller, A. (2016). Food classification and description. In B. Caballero, P. M. Finglas, & F. Toldrá (Eds.), *Encyclopedia of Food and Health* (pp. 1–6). Academic Press, Oxford.

Janowicz, K., Haller, A., Cox, S. J. D., Phuoc, D. L., & Lefrançois, M. (2018). SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *CoRR*, *abs/1805.09979*.

Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Kaboré, E. D. Y., Emonet, V., Graybeal, J., Laporte, M., Musen, M. A., Pesce, V., & Larmande, P. (2018). Agroportal: A vocabulary and ontology repository for agronomy. *Comput. Electron. Agric.*, *144*, 126–143. URL: `https://doi.org/10.1016/j.compag.2017.10.012`. doi:`10.1016/j.compag.2017.10.012`.

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). A decision support system for eco-efficient biorefinery process comparison using a semantic approach. *Comput. Electron. Agric.*, *127*, 351–367.

Madigan, D., Andersson, S. A., Perlman, M. D., & Volinsky, C. T. (1996). Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. *Communications in Statistics–Theory and Methods*, *25*, 2493–2519.

Mohammed, A.-W. (2016). Knowledge-oriented semantics modelling towards uncertainty reasoning. *SpringerPlus*, *5*.

Muljarto, A., Salmon, J., Neveu, P., Charnomordic, B., & Buche, P. (2014). Ontology-based model for food transformation processes - application to winemaking. In S. Closs, R. Studer, E. Garoufallou, & M. Sicilia (Eds.), *Metadata and Semantics Research - 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27-29, 2014. Proceedings* (pp. 329–343). Springer volume 478 of *Communications in Computer and Information Science*.

Munch, M., Dibie, J., Wuillemin, P., & Manfredotti, C. E. (2019a). Towards interactive causal relation discovery driven by an ontology. In *FLAIRS* (pp. 504–508).

Munch, M., Dibie-Barthélemy, J., Wuillemin, P., & Manfredotti, C. E. (2019b). Interactive causal discovery in knowledge graphs. In *Joint Proceedings of the 6th International Workshop on Dataset PROFIling and Search & the 1st Workshop on Semantic Explainability co-located with the 18th International Semantic Web Conference (ISWC 2019), Auckland, New Zealand, October 27, 2019* (pp. 78–93). CEUR-WS.org volume 2465 of *CEUR Workshop Proceedings*.

Munch, M., Wuillemin, P., Manfredotti, C. E., Dibie, J., & Dervaux, S. (2017). Learning probabilistic relational models using an ontology of transformation processes. In *On the Move to Meaningful Internet Systems. OTM 2017 Conferences - Confederated International Conferences: CoopIS, C&TC, and ODBASE 2017, Rhodes, Greece, October 23-27, 2017, Proceedings, Part II* (pp. 198–215). Springer volume 10574 of *Lecture Notes in Computer Science*.

Niculescu, R. S., Mitchell, T. M., & Rao, R. B. (2006). Bayesian network learning with parameter constraints. *J. Mach. Learn. Res.*, *7*, 1357—-1383.

Parviainen, P., & Koivisto, M. (2013). Finding optimal bayesian networks using precedence constraints. *The Journal of Machine Learning Research*, *14*, 1387–1415.

Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proc. of Cognitive Science Society (CSS-7)*.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. (2nd ed.). New York, USA: Cambridge University Press.

Poveda-Villalón, M., Fernández-Izquierdo, A., & García-Castro, R. (2019). Linked Open Terms (LOT) Methodology. Accessed: 2022-01-18.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*, 461 – 464. URL: `https://doi.org/10.1214/aos/1176344136`. doi:`10.1214/aos/1176344136`.

Shanmugam, K., Kocaoglu, M., Dimakis, A. G., & Vishwanath, S. (2015). Learning causal graphs with small interventions. *ArXiv*, *abs/1511.00041*.

Soldatova, L. N., Nadis, D., King, R. D., Basu, P. S., Haddi, E., Baumlé, V., Saunders, N. J., Marwan, W., & Rudkin, B. B. (2014). EXACT2: the semantics of biomedical protocols. *BMC Bioinform.*, *15*, S5.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. (2nd ed.). MIT press.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn Methodology for Ontology Engineering. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi (Eds.), *Ontology Engineering in a Networked World* (pp. 9–34). Springer.

URL: https://doi.org/10.1007/978-3-642-24794-1_2. doi:10.1007/978-3-642-24794-1_2.

Suzuki, J. (1996). Learning bayesian belief networks based on the minimum description length principle: An efficient algorithm using the b & b technique., . (pp. 462–470).

Truong, B. A., Lee, Y., & Lee, S. Y. (2005). A unified context model: Bringing probabilistic models to context ontology. *Embedded and Ubiquitous Computing – EUC 2005 Workshops*, (pp. 566–575).

Verny, L., Sella, N., Affeldt, S., Singh, P., & Isambert, H. (2017). Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, *13*, e1005662.

Zhang, S., Sun, Y., Peng, Y., & Wang, X. (2009). Bayesowl: A prototype system for uncertainty in semantic web. *Proceedings of the 2009 International Conference on Artificial Intelligence, ICAI 2009*, *2*, 678–684.

Zheng, H.-T., Kang, B.-Y., & Kim, H.-G. (2007). An ontology-based bayesian network approach for representing uncertainty in clinical practice guidelines. *CEUR Workshop Proceedings*, *327*.