A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models: application to processing of bio-composites for food packaging

Mélanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin, Helene Angellier-Coussy

A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models: application to processing of bio-composites for food packaging

Mélanie Münch[1][0000−0001−6704−1446], Patrice Buche[1,2][0000−0002−9134−5404], Cristina Manfredotti[3][0000−0003−4217−2591], Pierre-Henri Wuillemin[4][0000−0003−3691−4886], and Hélène Angellier-Coussy[1][0000−0001−5482−7095]

[1] IATE, U. Montpellier, INRAE, CIRAD, Montpellier SupAgro, Montpellier, France
[2] LIRMM, U. Montpellier, CNRS, INRIA GraphIK, Montpellier, France
[3] UMR MIA-Paris, AgroParisTech, INRAE, U. Paris-Saclay, Paris, France
[4] Sorbonne Universites, UPMC, U. Paris 06, CNRS UMR, LIP6, Paris, France
melanie.munch@gmail.com, patrice.buche@inrae.fr, cristina.manfredotti@agroparistech.fr, pierre-henri.wuillemin@lip6.fr

Abstract. Designing new processes for bio-based and biodegradable food packaging is an environmental and economic challenge. Due to the multiplicity of the parameters, such an issue requires an approach that proposes both (1) to integrate heterogeneous data sources and (2) to allow causal reasoning. In this article, we present POND (Process and observation ONtology Discovery), a workflow dedicated to answering expert queries on domains modeled by the Process and Observation Ontology (PO^2). The presentation is illustrated with a real-world application on bio-composites for food packaging to solve a reverse engineering problem, using a novel dataset composed of data from different projects.

Keywords: Ontology · Probabilistic model · Causality · Food packaging

1 Introduction

The massive amount of plastics used each year results in a constant accumulation of wastes in our environment, with harmful effects on our eco-systems and human health. Faced to the depletion of fossil resources and the increasing production of unrecovered organic residues (agricultural, urban, forestry and from agro-food industries), innovative technologies are developed for the production of bio-sourced, biodegradable and recyclable materials in order to increase the circularity of plastics. Among bio-polymers, 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 development of PHBV bio-composites loaded with lignocellulosic fillers is largely motivated by a decrease in PHBV's cost, an improvement

of the carbon footprint and a reduction of the global warming [6]. However, the augmentation of added lignocellulosic fibers has a negative impact over the bio-composite's brittleness and its process-ability. When developing bio-composites, a compromise must then be found between the maximum acceptable filler content, the filler size and the resulting properties. Yet, finding causal explanations for this compromise from data alone can be a challenging task. If previous works have suggested the use of interventions (i.e. changing a variable while keeping all other constant) to build causal models [23], in the case of bio-based food packaging, such interventions can become really time and money consuming. In this article, we present POND (PO^2 ONtology Discovery), a workflow dedicated to answering expert queries for domains modelled by the Process and Observation Ontology (PO^2) [17]. The main idea is to study Knowledge Bases (KBs) [11] using PO^2 to integrate expert knowledge into the learning of an extension of the Bayesian Networks (BNs), the Probabilistic Relational Model (PRM) [14]. While POND is able to answer a wide range of questions (qualitative and quantitative), in this article we focus on causal questions and illustrate the workflow with a real-world application on bio-based food packaging. Our original contributions are (1) the complete integration of PO^2 in a pipeline to answer expert queries, (2) a tool for answering causal assumptions that allows reverse engineering approaches and (3) a meta-analysis over multiple sources on bio-based packaging. Section 2 presents the background necessary for POND. It covers the PO^2 ontology, PRMs, as well as the combination of the two and causal discovery from data. Section 3 introduces our workflow and emphasizes its contributions to the state of the art on combining ontologies and probabilistic models and causal questions answering. Section 4 illustrates this workflow with a real-world application on bio-based packaging. This work has been defined in the framework of a regional (MALICE Languedoc-Roussillon) and two European (H2020 RESURBIS and NOAW) interdisciplinary projects involving computer scientists, data scientists and biomass processing experts for food and bio-based material production. MALICE project was the first to study several itineraries to produce composites using different biomass. It has been followed by RESURBIS (resp. NOAW) projects dedicated to urban (resp. agricultural) waste valorization.

2 Background

2.1 The Process And Observation Ontology

PO^2 is a generic process and observation ontology initially dedicated to food science [17], developed using the Scenario 6 of the NeON methodology [26], by re-engineering a first ontology for the eco-design of transformation processes [9]. It represents transformation processes by a set of experimental observations taken at different scales and links them on a detailed timeline. It has been recently used for bio-based products transformation process, especially food packaging design. Fig. 1 presents an overview of its different parts, it is described by 67 concepts and 79 relations. A **transformation process** is defined by a succession of **steps** inscribed in a **temporal entity**. To each step, multiple **components**

(which represent features of interest) can be added, themselves associated with different **results** and their corresponding units of measurements. PO^2 ontology version 2.0, implemented in OWL 2[5], is published on the AgroPortal ontology library[6], and is Creative Commons Attribution International (CC BY 4.0)[7].

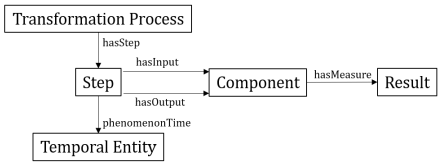


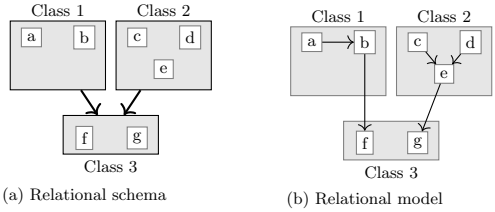
Fig. 1. Main parts of the PO^2 ontology.

2.2 Probabilistic Models: BN and PRM

A BN is the representation of a joint probability over a set of random variables that uses a directed acyclic graph (DAG) to encode probabilistic relations between variables. In our case, learning is done under causal constraints, which can be used to deduce causal knowledge through the essential graph (EG) [18], a semi-directed graph associated to the BN. Both the BN and its associated EG share the same skeleton, but the EG's edges' orientation depends on the BN's Markov equivalence class. A same edge's orientation for all equivalent BNs means that this orientation is necessary to keep the underlying probabilistic relations encoded in the graph: in this case, the edge is also oriented in the EG and is called an **essential arc**. Otherwise, it stays unoriented in the EG, meaning that its orientation does not modify the probabilistic relations encoded in the BN. In order to integrate expert knowledge under the form of causal constraints in the learning, we rely on PRMs, that extend BNs' representation with the oriented-object notion of classes and instantiations. PRMs [14] are defined by two parts: the **relational schema** RS (Fig. 2 (a)), that gives a qualitative description of the structure of the domain defining the classes and their attributes; and the **relational model** RM (Fig. 2 (b)), that contains the quantitative information given by the probability distribution over the different attributes. Classes in the RS are linked together by so-called **relational slots**, that indicates the direction of probabilistic links. Using these structural constraints, each class can then be learned like a BN[8], meaning they can be associated to an EG once instantiated.

Using constraints while learning BNs brings more accurate results, for parameters [7] or structure [8] learning. In case of smaller databases, constraining the learning can also greatly improve the accuracy of the model [21]. In this article we integrate expert knowledge as precedence constraints. Previous works already proposed methods for complete [5] or partial [22] node ordering. In our case we transcribe incomplete knowledge as a partial structural organization for the PRM's RS in order to discover new causal relations, as presented in [20].

[5] https://www.w3.org/TR/owl2-overview/
[6] http://agroportal.lirmm.fr/ontologies/PO2
[7] https://creativecommons.org/licenses/by/4.0/
[8] We use the classical statistical method Greedy Hill Climbing with a BIC Score.



(a) Relational schema (b) Relational model

Fig. 2. The high (a) and low (b) level structures of a PRM

2.3 Knowledge Discovery

Numerous works have proposed to use ontological knowledge in order to build probabilistic models and discover relations. For instance, different ontologies' expansions integrate probabilistic reasoning (such as BayesOWL [10], [28] or HyProb-Ontology [19]). These however do not allow the learning of relations. Other works directly uses the ontology's structure to build a BN, as for the objects properties that can be considered as probabilistic dependencies [13] or causal relations [1], which cannot however be applied with PO^2. Finally, some methods are tied down to specific cases, such as [2] that uses predefined templates to support medical diagnosis, which cannot be extended to other medical applications. While POND uses only PO^2, its complexity allows to deal with various tasks which gives it wider applications than a simple domain ontology.

For causal discovery, since correlation is not causation, the data set has to verify some conditions: no external factor (the **causal sufficiency** [25]); no missing or erroneous data, selection bias or deterministic cases [15]. In short, 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. Discovering causality from verified dataset can be done through independence tests between the variables [25], [27], but does not allow to introduce external constraints during the learning. Other works also proposed EGs to learn causal models: [16] presents two optimal strategies for suggesting interventions to learn causal models; [24] and [4] use an EG to build a causal BN (CBN) 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 encompassed in an ontology and a BN cannot be learned directly. Our goal is to use this knowledge to be as close as possible of a CBN, which is a BN whose relations' orientation translate a causal implication.

3 POND: PO² ONtology Discovery

We now present the POND workflow, whose aim is to integrate expert knowledge in order to query it. We focus here on how different sources can be studied in order to answer complex probabilistic and causal questions. A particular focus is cast on causal discovery and how it allows reverse engineering.

3.1 Knowledge Integration

Expert knowledge comes from: (1) experimental data, gathered from different sources (such as publications, books or data produced in different projects); and (2) direct interviews, where experts of a domain are solicited. This information is then structured under the PO^2 ontology. In our case, the interesting point is that all the data is now easily accessible thanks to its semantization. Once the data gathered and structured, the expert can express expert queries. Some can be answered through a simple query over the data described in the ontology (Competency Questions); others require a more in-depth analysis (Knowledge Questions, KQs). In this article, we will focus on **causal KQs** (cKQs), which can be formalized in two different ways. Given X_i and X_j groups of the domain's attributes:

cKQ_1 *Does* X_i *have a causal influence over* X_j*?*
cKQ_2 *What is the impact of* X_j *over* X_i*?*

Both illustrate the double reading offered by a CBN: while cKQ_1 focuses on the descriptive aspect, cKQ_2 allows to interrogate the nature of the relations between different variables. Once a cKQ expressed, we then build the probabilistic model. As seen in Section 2.3, we focus here on expressing the expert knowledge as a RS in order to guide the learning of the model. The originality of our approach is that this expression is done through two means:

1. **A mapping of the ontology's attributes in the RS.** Thanks to the common vocabulary defined by the PO^2 ontology, the expert can easily extract these attributes, even if they are measured in different contexts and depend on different sources of knowledge. For instance, a temperature might be measured at Step A with one source and at Step B with another. In this case, only the expert can tell whether these attributes are similar (i.e., if they can be compared) or not. With PO^2's semantic, the expert can thus select the attributes that are interesting to study, by specifying the process, step and component that lead to the interesting result (i.e., the datatype property which owns the value). This combination of results composes the BN's learning database.
2. **A definition of the precedence constraints.** Precedence constraints are possible orientations between the attributes encoded in the RS: if a relation is learned between two attributes linked by such an orientation, the learnt relation has to be oriented following it. These precedence constraints can either be deduced from the temporal information of PO^2 (a change of an attribute at time t may have an influence over an attribute at time $t+n$, but not at time t-n), or given by the expert according to their own knowledge ("I know that X_1 may have an influence over X_2").

Our contribution in this section is the automation of this knowledge integration in a workflow: thanks to PO^2, any transformation process can be easily integrated into a RS, using only a vocabulary specific of the studied domain.

3.2 Causal Discovery

Once the RS defined, a PRM can be learned and then instantiated as a BN. Since this is done under causal constraints, we can use the EG to deduce causality [20]. Indeed, the resulting model can be seen as the intersection of all the models constrained by the dataset used for the learning (expressed in the EG) and all the models constrained by the expert knowledge (expressed in the RS). Although it is not usually enough to learn a CBN, the EG's essential arcs can be used to complement expert knowledge. The causal validation is done as follows:

- If a relation is learned between two variables with an expert precedence constraint, then the causality is validated by the expert's knowledge.
- If a learned relation is an essential arc on the EG, then the causality is validated by the EG. This is the case even if no precedence constraint has been placed between those attributes.
- If a relation is learned, but is neither an essential arc nor part of a precedence constraint, then it is impossible to deduce causality.

Even if a complete CBN is not learned, this causal discovery has two goals:

- **Helping the expert criticize.** Since we aim to learn a real-world model, the evaluation of its performances cannot be done directly. However, by presenting the learned causal relations to the expert, we give them a tool to criticize and question it. An example of this critic is given in section 4.3.
- **Answering the cKQs.** cKQs depend on causal discovery to be answered: cKQ_1 directly requires the presence (or absence) of causal relations and in order to express the interactions questioned by cKQ_2, we need first to define the causality between the studied variables.

3.3 Causal Inferences

While what was explained in the previous section is enough to answer cKQ_1, answering cKQ_2 requires a more in-depth analysis. To illustrate this, we consider the CBN presented in Fig.3 as the result of a causal validation, and the following cKQ_{ex}: *"Which intervention should I do on the accessible variables to maximize the variable E?"*, which is a sequence of the cKQ_1 (*"Which variables have a impact over E?"*) and the cKQ_2 (*"What is the influence of these variables?"*).

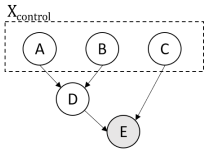


Fig. 3. Example of a CBN. The set $X_{control}$ represents the control variables, meaning the ones on which the expert can intervene; E is the target variable.

In order to answer cKQ_{ex}, we first need to assess which variables in $X_{control}$ (the set of variables on which the expert can intervene) are necessary. In our

case, we see that the direct parents of E are D and C. However, D is not in $X_{control}$, so we need to look at its own parents, which are A and B. Since they both belong to $X_{control}$, then in order to answer cKQ_{ex}, we define $X_{inter} = \{A, B, C\}$. Because we consider a CBN, then intervening on X_{inter} will have an effect over the target E. In practice, for each possible combination of values of X_{inter}, we can predict the values of E and their associated probability, which constitute a base of possible scenarios. In order to sort these, the expert expresses their own criteria of acceptability, as "which values are better for the target variable", or "which conditions should apply on X_{inter}". These criteria can be of two kinds:

- **Hard Criteria.** Some values or combinations of values are impossible to obtain: these scenarios are automatically discarded. For instance, the expert might wish that the sum of the values from X_{inter} does not exceed a certain value; or they might want to exclude some values for E (in our case, the goal is to maximize E: thus, it is not interesting to consider the lowest values).
- **Soft Criteria.** In this case, the expert needs to sort their preferences regarding the context. Maybe having a high value for E is not interesting if A also needs to be high; or a lower value for E with a higher probability might be more interesting than a better scenario with less chances of happening.

Defining these criteria helps the expert to select an answer corresponding to their need. As seen in Sec. 4.3, this can be used to do reverse engineering, whose goal is to understand how a system works through deductive reasoning. Sec. 4.3 shows an example where we formulate the composition of an optimal biomass.

4 Application to Bio-composites Packaging Materials

Given the context of bio-packaging, we define cKQ_{bio}: *"Which filler allows to optimise the packaging's tensile properties?"*.

4.1 Knowledge Base Presentation

Data was collected from four projects focused on the development of PHBV-based bio-composites using lignocellulosic fillers (LFs) stemming from organic waste streams, e.g. crop residues (*Chercheur d'avenir region Languedoc-Roussillon MALICE* and *NoAW*), agro-food by-products (*FP7 EcoBioCAP*) or urban waste (*H2020 Resurbis*). LFs were obtained by dry fractionation of the raw biomass. Pure cellulose fibers were also used as reference, representing in the end a database of 85 samples with 15 attributes.

4.2 Expert Integration

Integrating expert knowledge requires the expert to map from the knowledge base to the RS the attributes relevant for the cKQ, and to organize their potential precedence constraints. In this section, we present the main results used to learn our final model, as well as an example of the integration of some expert critics.

Attributes selection.[9] The expert describes LFs by three main categories: biochemical composition with the plants' main organic (**cellulose**, **hemicellulose**, **lignin**) and inorganic (**ash**) compounds; apparent median diameter (**D50**); **filler content**. Tensile parameters were determined from stress-strain curves obtained by tensile tests performed until the break of materials. The **Young's modulus** (slope of the initial section of the curve), **stress at break** (stress value at moment of material fracture) and **strain at break** (elongation value at moment of material fracture) respectively characterize the stiffness, the resistance and the ductility of the material. While these are enough to consider cKQ_{bio}, the expert helped us determine three other categories, in order to offer a better overview for the expert feedback: **permeability** (to water vapour), thermal properties (**crystallization** and **melting** temperatures) and thermal degradation (**onset** and **peak** temperature). Discretization is important, as it can influence the learning of the different relations and may be subject to change depending on the feedback from the expert. Table 1 presents an excerpt of it, where control variables are evenly distributed, while others follow a distribution chosen by the expert.

Lignin	]0;19.4] *(32)*	]19.4;26.4] *(30)*	]26.4;49[*(23)*				
Filler Content	]2;4] *(10)*	]4;11] *(34)*	]11;21] *(22)*	]21;50[*(19)*			
Strain at Break	]0.2;0.5] *(19)*	]0.5;0.8] *(44)*	]0.8;1] *(15)*	]1;1.07[*(3)*			

Table 1. Example of the discretization used for some variables *(number of examples)*.

Precedence Constraints Definition. The expert defines two precedence constraints that may be refined after each iteration.

- Between the filler variables and the package's characteristics. We consider the first as control variables, whose values may have an impact over the final result. We create two classes in the RS, with a relational slot from the control variable's class towards the package's characteristic's class.
- Between the different package's characteristics. They cannot influence each other (e.g. the tensile attributes have no influence over the thermodynamic ones). As a consequence, we compartmentalise the RS characteristic's class into different separated sub-classes, such that they have no relational slot except the one from the control variable class.

Expert Feedback. Once a model is learned, discussion with an expert is required to criticize both (1) the learned relations and (2) the probabilistic dependencies. For example, in Fig. 4, the expert mentioned that the **crystallization** temperature could not be explained by the **melting** parameter, and that the learned relation translates a correlation, not a causation. As a consequence, we create a constraint that prevents the learning of this link. Finally, **strain at break** was not expected to not be explained by any parameter, which suggested to try a new discretization to better represent the variable. The expert is also useful to explain the lacks of knowledge. Regarding the **melting** temperature, this model highlights (through near-zero probabilities) that if **content** $\in]21;50[$, then **melting** $\notin]1;1.02[$. This was fully expected since the melting temperature is not supposed to increase when adding LFs.

[9] For the rest of the article, all attributes represented in the model are **bolded**.

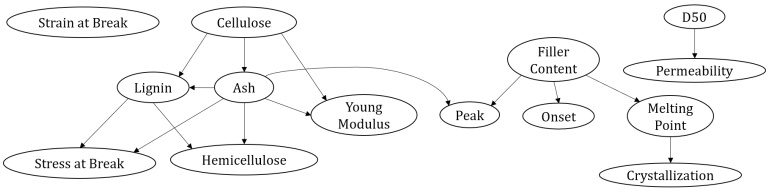


Fig. 4. Model learned after one iteration.

4.3 Knowledge Question Answering

We now consider the CBN accepted by the expert, presented in Fig. 5. For the sake of the example, we present a simplified version where all non-relevant variables were removed. cKQ_{bio} addresses two possible interventions for improving the three considered tensile properties: (1) **filler content** and (2) **LF**.

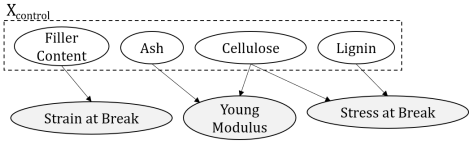


Fig. 5. Extract of the BN selected for Biomass Discovery. Since all relations are influenced by precedence constraints, we consider this as a CBN validated by the expert.

Finding Optimal Content. According to Fig. 5, **Filler Content** has a causal influence only on **Strain at Break**. Depending on the expert's criteria, multiple readings of the conditional probability table (Table. 2) are possible:

- When aiming for the highest value possible for **Strain at Break** (]1;1.07]), the probabilities are almost zero. Thus, it cannot realistically be satisfied.
- With a hard criteria aiming for the second highest value of **Strain at Break**, a **content** of]2;4] could be considered, as it guarantees a probability of 0.3963 to obtain the second best value (]0.8;1]).
- In the case of an industrial process, however, the expert might want to place a hard criteria for a reasonable probability of success. In this case, a **content** of]4:11] should be applied, since it guaranties a probability of success of 0.7.

	Strain at Break			
	]0.24;0.5]	]0.5;0.8]	]0.8;1]	]1;1.07]
Filler Content				
]2;4]	0.0061	0.5915	***0.3963***	0.0061
]4;11]	0.002	***0.7060***	0.2260	***0.0660***
]11;21]	0.3623	0.4972	0.0927	0.0478
]21:50]	***0.6062***	0.2774	0.113	0.0034

Table 2. Conditional probabilities of **Strain at Break** (***maximum likelihood***).

Proposing new LF. According to the BN presented in Fig. 5, **Young's Modulus** and **Stress at Break** depend on components of the biomass. We first define some criteria of acceptability:

- **Hard criteria HC_1.** The sum of the **ash**, **cellulose**, and **lignin** must not exceed 100 (i.e. the biomass must be possible). We fix HC_1 such that, given $x \in \{$**Ash**, **Cellulose**, **Lignin**$\}$ and its interval $[x_{min}; x_{max}]$, we have $\sum_x x_{min} < 100$.

- **Hard criteria HC_2.** We want the target variables within interesting range of values, and fix **Stress At Break** $> 0.8 \cap$ **Young Modulus** > 0.8.
- **Hard criteria HC_3.** The probability of success must be higher than 0.25.
- **Soft criteria SC_1.** When no corresponding biomass is found, we allow the system to look for similar ones, that can be considered close to the one we are looking for. Given a biomass m in AtWeb, its composition x_m and a target interval $[x_{min};\ x_{max}]$ (with $x \in \{$**Ash**, **Cellulose**, **Lignin**$\}$), we define a score $S_m = \sum_x \sigma(m, x)$
  with $\sigma(m,x) = \begin{cases} 0 \text{ if } x_m \in [x_{min}; x_{max}]; \\ min(abs(x_m - x_{min}), abs(x_m - x_{max})) \text{ otherwise.} \end{cases}$
  The lower S_m is, the closer the biomass is to our recommendation.

In order to suggest new biomasses for packaging composite making, @Web RDF database [3] including experimental data about biomass deconstruction [12] has been queried using these criteria, which returned five solutions (Table 3 presents the first three). Each of these scenarios assesses the probability of obtaining a value over 0.8 for the tensile properties. The most probable one ($p = 0.41$) is not an exact match; however, the closest match, the rice husk, has an S-score of 0.73, meaning it is really similar to the scenario's recommendations. This corroborates with the second scenario, which also recommends the rice husk with a slightly lower probability of outcome. The last scenario, finally, proposes the pine bark, with a S-score of 5.24 (due to the pine bark's **ash** value of 1.44). It is important to note that a limit of this model is tied to the discretization required by BN learning. When dealing with values close to the border of the interval, predicting the result is more difficult. Moreover, Table 1 shows that some categories are underrepresented compared to the others (e.g. **Strain at Break** $\in$]1;1.07]). If this choice of discretization bears a meaning for the domain, it however introduces bias: some categories may artificially have a bigger weight than the others during the learning only because they do not have enough samples. That is why the database used for the learning must be really representative, to allow a smoother discretization which would prevent this edge effect.

p	0.41	p	0.40	p	0.28
Ash	[6.7;24.7]	Ash	[6.7;24.7]	Ash	[6.7;24.7]
Cellulose	[25.6;33]	Cellulose	[10.9;25.6]	Cellulose	[10.86;25.59]
Lignin	[26.4; 49]	Lignin	[19.4; 26.4]	Lignin	[19.4; 26.4]
Exact Match	$\emptyset$	**Exact Match**	Rice Husk	**Exact Match**	$\emptyset$
Close Match	Rice Husk	**Close Match**	$\emptyset$	**Close Match**	Pine Bark
$S_{RiceHusk}$	0.73	S	$\emptyset$	$S_{PineBark}$	5.24

Table 3. Results of Biomass Querying with respect to HC_1, HC_2, HC_3 and SC_1. When no exact result, a S-score was calculated to find the closest match.

4.4 Conclusion

In this paper we have presented POND, a complete workflow dedicated to answer EQs over processes represented by the PO^2 ontology. We focused on causal discovery aspects and illustrated it with a real-world example, the bio-packaging transformation process. Thanks to the use of the ontology, this workflow allows the expert to easily handle the knowledge integration part and to add more

knowledge under the form of precedence constraints. During the answering, they can also express criteria of acceptability to elect the best answer for their needs. As in all causal discovery contexts, multiple conditions must be verified in order to be accepted, as described in Section 2.3. This also requires the expert to be trustworthy, both for the constraints' definition and the model verification. Finally, as presented in the example, a database too sparse for the learning could lead to questionable discretization that could be difficult to interpret. Future works will look into the use of the answers to assess the quality of the current KB and see how it can be used either to suggest correction for the current base, or generation of new data to fulfill knowledge holes. Another interesting task would be to address the dedication of POND to the PO^2 ontology, which represents a limit; while the method should work in theory with any other semantic structuration of the data, it needs to be reworked to be adapted.

Acknowledgement

We would like to thank Claire Mayer (PhyProDiv Team, INRAE IATE) who provided data for the biomass discovery aspect. Our work has been partially financed by the French national research agency ANR in the framework of D2KAB (ANR-18-CE23-0017) and DataSusFood (ANR-19-DATA-0016) projects.

References

1. Ben Messaoud, M., Leray, P., Ben Amor, N.: Semcado: A serendipitous strategy for learning causal bayesian networks using ontologies. Symbolic and Quantitative Approaches to Reasoning with Uncertainty pp. 182–193 (2011)
2. Bucci, G., Sandrucci, V., Vicario, E.: Ontologies and bayesian networks in medical diagnosis. HICSS pp. 1–8 (2011)
3. Buche, P., Dibie-Barthelemy, J., Ibanescu, L.L., Soler, L.: Fuzzy Web Data Tables Integration Guided by a Termino-Ontological Resource. IEEE Transactions on Knowledge and Data Engineering **25**(4), 805–819 (2013)
4. Castelletti, F., Consonni, G.: Discovering causal structures in bayesian gaussian directed acyclic graph models. Journal of the Royal Statistical Society Series A, Royal Statistical Society **183**, 1727–1745 (2020)
5. Cooper, G.F., Herskovits, E.: A bayesian method for the induction of probabilistic networks from data. Machine Learning **9**(4), 309–347 (1992)
6. David, G., Croxatto Vega, G., Sohn, J., Nilsson, A.E., Hélias, A., Gontard, N., Angellier-Coussy, H.: 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 (2020)
7. De Campos, C.P., Ji, Q.: Improving bayesian network parameter learning using constraints. In: ICPR. pp. 1–4 (2008)
8. De Campos, C., Zhi, Z., Ji, Q.: Structure learning of bayesian networks using constraints. In: ICML. pp. 113–120 (2009)
9. Dibie, J., Dervaux, S., Doriot, E., Ibanescu, L., Pénicaud, C.: $[MS]^2O$ - A multiscale and multi-step ontology for transformation processes: Application to microorganisms. In: ICSS. pp. 163–176 (2016)

10. Ding, Z., Peng, Y., Pan, R.: BayesOWL: Uncertainty Modeling in Semantic Web Ontologies, pp. 3–29. Springer Berlin Heidelberg (2006)
11. Ehrlinger, L., Wöß, W.: Towards a definition of knowledge graphs. In: SEMANTiCS (Posters, Demos, SuCCESS) (2016)
12. Fabre, C., Buche, P., Rouau, X., Mayer-Laigle, C.: Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. Data in Brief **33** (2020)
13. Fenz, S.: Exploiting experts' knowledge for structure learning of bayesian networks. Data And Knowledge Engineering **73**, 73 – 88 (2012)
14. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: IJCAI. p. 1300–1307. Morgan Kaufmann Publishers Inc. (1999)
15. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Frontiers in Genetics **10**, 524 (2019)
16. Hauser, A., Bühlmann, P.: Two optimal strategies for active learning of causal models from interventional data. International Journal of Approximate Reasoning pp. 926—-939 (2014)
17. Ibanescu, L., Dibie, J., Dervaux, S., Guichard, E., Raad, J.: Po2- a process and observation ontology in food science. application to dairy gels. Metadata and Semantics Research pp. 155–165 (2016)
18. Madigan, D., Andersson, S.A., Perlman, M.D., Volinsky, C.T.: Bayesian model averaging and model selection for markov equivalence classes of acyclic digraphs. Communications in Statistics–Theory and Methods **25**(11), 2493–2519 (1996)
19. Mohammed, A.W.: Knowledge-oriented semantics modelling towards uncertainty reasoning. SpringerPlus **5** (2016)
20. Munch, M., Dibie, J., Wuillemin, P., Manfredotti, C.E.: Towards interactive causal relation discovery driven by an ontology. In: FLAIRS. pp. 504–508 (2019)
21. Munch, M., Wuillemin, P.H., Manfredotti, C., Dibie, J., Dervaux, S.: Learning probabilistic relational models using an ontology of transformation processes. In: OTM 2017 Conferences. pp. 198–215 (2017)
22. Parviainen, P., Koivisto, M.: Finding optimal bayesian networks using precedence constraints. Journal of Machine Learning Research **14**, 1387–1415 (2013)
23. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edn. (2009)
24. Shanmugam, K., Kocaoglu, M., Dimakis, A.G., Vishwanath, S.: Learning causal graphs with small interventions. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 28. Curran Associates, Inc. (2015)
25. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search. MIT press, 2nd edn. (2000)
26. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The neon methodology for ontology engineering. In: Suárez-Figueroa, M.C., Gómez-Pérez, A., Motta, E., Gangemi, A. (eds.) Ontology Engineering in a Networked World, pp. 9–34. Springer (2012)
27. Verny, L., Sella, N., Affeldt, S., Singh, P., Isambert, H.: Learning causal networks with latent variables from multivariate information in genomic data. PLOS Computational Biology **13** (2017)
28. Zhang, S., Sun, Y., Peng, Y., Wang, X.: Bayesowl: A prototype system for uncertainty in semantic web. ICAI **2**, 678–684 (2009)