DCPO: The dairy cattle performance ontology, a tool for domain modelling and data analytics

V. Fuentes[12], T. Martin[1], P. Valtchev[1], A. B. Diallo[1,2], R. Lacroix[3] and M. Leduc[2]

[1]CRIA, Département d'informatique, UQÀM, Montréal (QC), Canada
Corresponding Author: diallo.abdoulaye@uqam.ca
[2]LACIM, Département d'informatique, UQÀM, Montréal (QC), Canada
[3]Lactanet, Sainte-Anne-de-Bellevue, Canada

Abstract

Dairy farming is being intensively computerized, whereby the goal is to use the recorded data to optimize production processes. This requires extensive analytics, which needs a good understanding of the data. It is also necessary that the datasets be federated to be able to get an integrated view. Although conventional database tools are helpful in that process, it is believed that linked data and ontologies can provide seamless integration of different sources while providing a semantic layer allowing deeper introspection of data. The objective was to build an ontology to provide such a semantic layer to dairy herd improvement (DHI) data.

A large dataset of milk production data was provided by Lactanet, Canadian Network for Dairy Excellence. This data is typically heterogeneous, i.e., covering partially or thoroughly health, nutrition, yield, and genetics. It also possesses a complex structure, with a large variety of data for a unique animal, dispersed in many records and multiple tables. A dedicated domain ontology, referred to as the Dairy Cattle Performance Ontology (DCPO), was built from a semantic analysis of the datasets. The initial core set of entities was determined using the definitions and minimal attribute sets for traits provided by ICAR guidelines and CDN documents. This core was gradually enriched with lower-level entities and aligned to more abstract concepts from the Basic Formal Ontology (BFO) to provide a foundational theory. The process was validated by domain experts. DCPO provides a rich and extensible data schema, a vocabulary based on international standards to support stakeholder collaboration. It federates external data sources and provides a semantic interface to query the obtained integrated linked data. Finally, DCPO underlies a knowledge base supporting analytics and decision making. Preliminary evaluation followed a query-based approach: SPARQL queries were designed reflecting typical questions experts might ask to assess the practical usability of DCPO.

Mining structural regularities, or patterns, in data may lead an expert to discover unknown phenomena or to confirm an already formulated hypothesis. The benefit of using DCPO as vocabulary for patterns is to enable seemingly unrelated yet isomorphic sub-graphs in the data with diverging vertex and edge labels, to become identical once their labels are generalized to DCPO classes and properties. Key benefit thereof was the patterns were described using the domain expert language to increase their interpretability. Next, we plan to use the ontology to support the deep learning-based inference of predictive models for milk production.

Keywords: Precision agriculture, dairy farming, domain ontologies, knowledge discovery from data, graph mining.

Introduction

Agriculture 4.0[1] refers to future trends helping the sector face the main challenges pertaining to the demands of the future. Precision farming, in particular, is about improving the overall farming process through in-depth analysis of its various aspects as reflected in their historical data generated by farming devices, produce/crop processing entities, regulatory bodies, etc. This requires all the stakeholders (e.g., producers, managers, analysts, consultants, etc.) to work together to leverage available data as a competitive advantage.

A typical approach is the design of machine learning or data mining-based analytical tools to, *inter alia*, predict outcomes in daily-life situations the stakeholders face or to detect major trends and/or exceptional events in the data. As living beings are involved, data are typically heterogeneous and complexly structured: They may cover such aspects as the well-being and health issues for farming animals, nutrition, yield, genetics, etc. Inner structure, e.g., time series, and inter-record relations, e.g., animal pedigree, would also appear in the data.

Constituting such complexly structured datasets requires a significant data-modelling effort. Moreover, as ever more aspects of the farming process get computerized, extensibility to further datasets is often a prime concern. This motivates a full-scale domain modelling in the form of a dedicated domain ontology (DO). DOs have a wide range of benefits beyond mere rich/extensible data schema. For instance, they provide a standardized vocabulary to support stakeholder collaboration while representing a centralized repository for domain expertise, thus enabling the design of decision-support systems for various domain tasks [1]. We present here the design of our *dairy cattle performance ontology* (DCPO), its current state and intended usage. The remainder of the paper is as follows: Section 2 presents our motivations while section 3 lists relevant prior work. Next, section 4 details our iterative modeling process and our tool set. Finally, section 5 concludes.

Motivation

Lactanet[2], is the dairy production centre of expertise covering the province of Quebec and the Atlantic regions of Canada. Lactanet's accumulated data about dairy production and milk control describes 6,670 herds and 1.5M cows. Key concepts reflected in the data include milk control samplings and the associated laboratory-based analyses that estimate the principal milk components: fat, protein, milk urea nitrogen, etc. Overall, the records provided by Lactanet amount to 3+ billion data end points. This huge dataset hides potentially meaningful concepts, e.g., unproductive cows admitting improvement vs those to quickly sell, and behavioral patterns for cows or farmers, that need to be uncovered. In order to allow richly structured heterogeneous datasets to be: (1) properly built and (2) analyzed to yield meaningful and intelligible patterns, we decided to design a DO. A number of our dairy analytical tools are symbolic-level, including graph mining methods whose cornerstone is a DO-powered generalized pattern miner. Additionally, a set of predictive models exploiting deep neural net architectures have been designed targeting a variety of yield metrics such as milk production and overall cost [2]. The way these can benefit from the ontology and the graph mining tools' output is currently under investigation.

[1] https://www.worldgovernmentsummit.org/api/publications/document?id=95df8ac4-e97c-6578-b2f8-ff0000a7ddb6

[2] https://lactanet.ca/en/home

Related work

Several ontological sources have been developed about dairy production. The *Animal Trait Ontology for Livestock* (ATOL, https://www.ebi.ac.uk/ols/ontologies/atol) models phenotypical animal traits. These are represented from an environment-aware and animal breeding-driven point of view. A *Common Dairy Ontology* (CDO) [3], has been designed towards assisting on farming decision making and semantic alignment (www.smartdairyfarming.nl).Yet CDO is primarily focused on sensor data and lacks a transverse view of the domain. *AgroRDF* [4] is a data exchange standard designed for agro-industrial purposes and built with semantic technologies. However, it lacks a unifying broader framework able to precisely describe the dairy domain. The *agriOpenLink* [5] system provides open interfaces and linked services to enable the development of new processes with a plug-and-play architecture. The *Dairy Farming Ontology* (DFO) is among the many created within the agriOpenLink project. Albeit strongly appealing for our own goals, it is not publicly available. The FAO (Food and Agriculture Organization of the United Nations) project develops agricultural standards such as *AgroVOC* vocabulary [6]. While it covers a wide range of subjects, it lacks middle-level concepts involved in dairy production, hence it is too generic for our needs.

In the recent past, DOs have been used to support a semantically rich data mining process. Indeed, they expose domain knowledge to machine processing while providing a rich vocabulary that is easily intelligible for domain experts [7]. Pattern mining [8], [9] aims at discovering recurrent data fragments in a dataset that might represent potentially useful trends and regularities (combinations of descriptors). Depending on data record topology and how much thereof is preserved in the patterns, various flavors of patterns have been studied, from itemsets (sets of products) to sequences to graphs. Independently, *generalized* patterns [10] have been introduced to deal with cases where abstracting from concrete data items (e.g. *Corona virus* instead of *SARS CoV 2* can bring insights absent in the ground level of data records. Generalized patterns are defined on top of an item taxonomy.

Graphs are among the most challenging pattern formats and adding a DO on top of their vertex and edge labels further compounds the issue. Partial solutions to the graph mining with a DO problem were investigated in [11]–[13]. Both [12] and [13] under-exploit the ontological structure by focusing only on parts of it (object properties and classes, respectively). In comparison, our DO is intended to support abstraction on edges as well, e.g. use *parent* property in patterns to match the *dam* property (female parent of a bovine) in data. In [11], abstraction from both vertices and edges was formalized, yet for graphs built around a vertex sequence which largely eases the mining task. In contrast, we deal with unrestricted graphs.

Finally, the problem of feeding the knowledge from a DO into a neural learning process was approached in [14] with class-embedding-based techniques. Prior studies have investigated mimicking the ontological structure by the neural network architecture [15]. Unlike these, we rely on discovered graphs patterns for data augmentation [16].

Building the dairy cattle performance ontology

We were provided with several non-ontological resources such as datasets of various provenance and coverage pertaining to dairy production, together with their data dictionaries. Additionally, we followed the International Committee for Animal Recording (ICAR, www.icar.org) guidelines and the publication of dairy cattle genetic evaluations in Canada provided by the Canadian Dairy Network (CDN, https://www.cdn.ca), consisting of publicly available data dictionaries. Starting from these resources, we applied an iterative modeling process inspired by the Ontology Summit 2013 Communiqué's life cycle (://ontolog.cim3.net/OntologySummit/2013/communique.html. Below, we describe its main steps and outcomes.

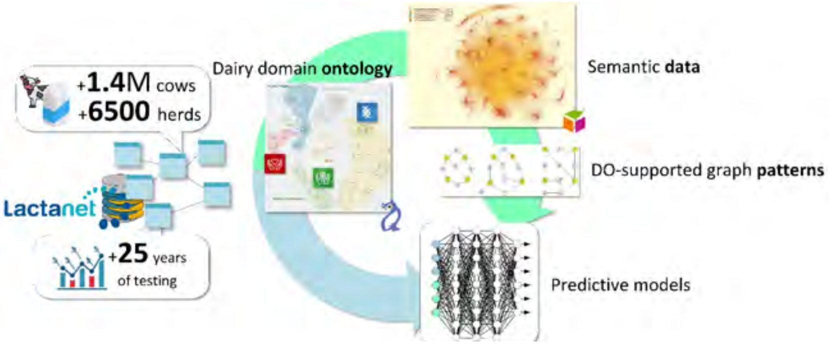


Figure 1. Global project architecture.

Requirements

Our initial focus was on purpose and scope of the ontology. Figure 1 depicts the global architecture of the information system the ontology is intended to support.

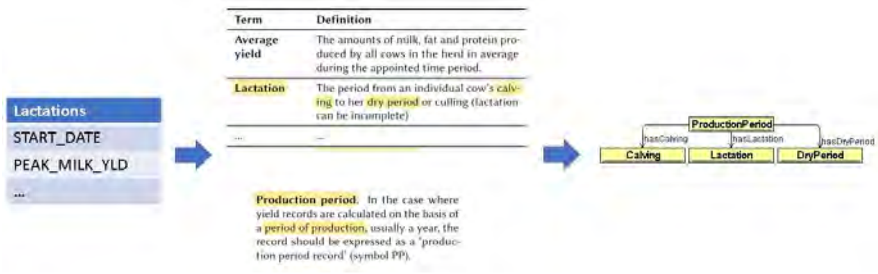
The DCPO, in the center, is intended to support three use cases: (1) a federated schema for external data sources on the left, (2) a semantic interface to cross-domain querying the linked data produced from the integration of external and internal sources, and (3) a knowledge base for graph mining algorithms on the right side. The produced generalized graph patterns are injected into further machine learning tools as additional ontology-based features to make their results more intelligible. As the results of this ontology supported system are intended to be used by anyone in the dairy community, it is important the compliance to international standards in the field, in particular to the ICAR guidelines and CDN’s genetic vocabularies.

Scope

Precision dairy farming is about optimizing dairy cattle performance indicators that reflect complementary aspects of the dairy process. In the long term, DCPO will embody knowledge about six perspectives of the dairy process: breeding (pedigree), genetics, production and milk quality control, environment, health and nutrition. Currently, it encompasses only the first four, corresponding to the datasets and data dictionaries made available as departing point.

Ontological analysis and design

The ontological analysis for DCPO has been guided by domain experts, experienced data scientist, and the available structured description of the dairy data recording procedures within the ICAR documentation. This documentation was informative enough as to provide a core set of terms that was gradually enriched with lower-level concepts and properties and aligned to an upper ontology to provide a foundational theory.



Term	Definition
Average yield	The amounts of milk, fat and protein produced by all cows in the herd in average during the appointed time period.
Lactation	The period from an individual cow's calving to her dry period or culling (lactation can be incomplete)
...	—

Production period. In the case where yield records are calculated on the basis of a period of production, usually a year, the record should be expressed as a 'production period record' (symbol PP).

Figure 2. Bottom-up analysis: From dictionary names, via ICAR terms, to DCPO entities.

To take advantage of the available resources mentioned before, a bottom-up approach was performed. To identify the key entities of the ontology, we extracted the names of our datasets and their columns from available data dictionaries and matched them to the terms defined in the ICAR documents, so that we could link them to the standardized dairy domain terminology, Figure 2 illustrates this process. On the left, the candidate term *Lactation* is retrieved from the dataset name, and it's matched to the terms defined in the ICAR document, where an occurrence is found. In the matched definition, the related candidate terms/phrases *Calving* and *DryPeriod* are retrieved. Additional examination of the document identifies another candidate term *ProductionPeriod* and its relationship to the other terms are inferred (e.g. *hasLactation*).

In defining hierarchies (i.e., classes and properties) we usually provide an abstract level to factor out the common characteristics of the elements in a particular module, and one or more specialization levels below which inherit and refine these characteristics. This facilitates the management of the overall ontology architecture and inter-module connections, its extension, readability, and better grouping of similar entities. In some cases, the generalization process leads to the finding of *ontology patterns* (OP) that can then be reused across the ontology to provide modularity. Moreover, this decomposition of the ontology in abstraction layers allow the discovery of *generalized graph patterns* (GGP) from data, as we will see in the next section.

One such OP is the *ascertainment pattern* shown in Figure 3 (left). In detail, a target, *Thing*, undergoes an assessment procedure of some kind or *Ascertainment*, about some *DeterminableQuality* of the target and it is quantified by some measure or *DeterminateQuality*. This OP abstracts different ways of acquiring certain knowledge concerning the target entity. Under the umbrella of this OP, one finds such dairy farming activities as milk composition tests, genetic evaluations, and cow conformation scoring, to name a few. For instance, genetic evaluation is depicted on the right of Figure 3.

In searching for abstractions and OPs, we combine the bottom-up strategy of generalizing from concrete entities with the top-down strategy of making them specializations of a foundational ontology, the *Basic Formal Ontology* (BFO) from the OBO Foundry in our case. This greatly simplifies the integration of the two ontologies as the specialization approach gradually refactors the DCPO using BFO as a design guide, trying to align our entities to entities in the upper ontology. This has the effect of forcing our design to comply to the upper ontology, and thus absorb its principles. As an example, a genetic trait is any measurable characteristic of a cow that is heritable with some probability. Using the bottom-up strategy, we found a hierarchy of trait classes associated with concrete measures. From the top-down perspective

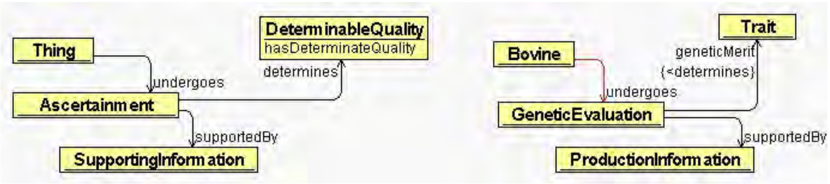


Figure 3. The OP Ascertainment (left) and its instantiation as GeneticEvaluation (right).

we understand that traits are *BFO*:*Quality* specializations. So, we created the classes *DeterminableQuality* and *DeterminateQuality* for general use in our patterns, which are both specializations of *BFO*:*Quality* and generalizations of our concrete classes for traits and measures correspondingly (see Figure 3).

This strategy enabled the rapid design of a coarse first model made of candidate classes and properties. We chose OWL since it is a standard Semantic Web technology built on top of RDF, a data format designed for interoperability, and provides valuable inference capabilities allowing ontology consistency to be checked with a reasoner, thus greatly reducing the production effort for the formal ontology artifact. Domain experts challenged this first model/design. This strategy enabled an analysis/design process following an iterative feedback loop (refining, updating, and adding new entities) with domain experts regularly challenging the latest changes.

Ontology description

The ontology has been modularized according to the different aspects of the dairy process covered at this stage: core, production and quality control, testing, breeding, and genetic evaluation. In the following paragraphs we summarized the ontology description as depicted in Figure **4**. Notice the use of italics to highlight ontology entities where classes begin with an uppercase letter and properties with a lowercase one.

At the core of the ontology, the central entity *Bovine*, factors out common characteristics of main actors: *Cow* and *Bull*, regardless of their particular role in the process or their life stage, allowing these concrete specializations to refine a common base by inheritance. *Bovine* is derived from *Animal*, used to enable extensions of the DCPO to other dairy species. A *parent* property and its specializations *femaleParent* and *maleParent* are defined on *Animal* to allow the construction of a parentage graph tracking the pedigree of each animal, with further specializations *dam* and *sire* for the cows and bulls involved in breedings, respectively.

The productive periods of a cow have three main stages: *Calving*, representing the birth of a new calf; *Lactation*, the milk production periods the cow has went through and *DryPeriod*, the time the cow is not producing milk. During *Lactation*, the *Milking* of cows undergoes *QualityControl* whose instances represent the different milk quality checkpoints performed during lactation. Quality control performs *MilkSampling* to produce a *MilkSample* that *isAnalyzedBy* a *CompositionTest*. During *MilkSampling* a *QuantificationTest* is performed to measure the milk yield . A *Breeding* between a *breedingDam* and a *breedingSire engenders* a new *Calving* producing a newborn *Bovine*. Each *Bovine*, undergoes a *GeneticEvaluation* that determines the *geneticMerit*

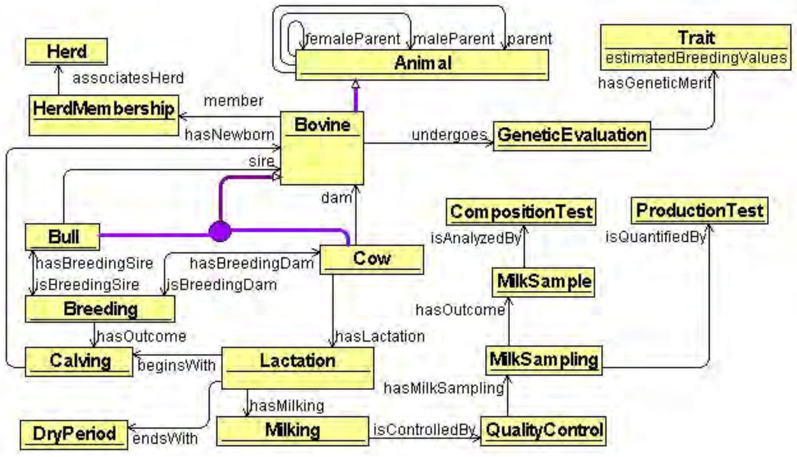


Figure 4. Dairy ontology and its modular desig.

of the animal on several *Trait*s, to assess its value (a full description is available through CDN). Finally, a *Herd* entity is associated with the concept of *HerdMembership*, representing the fact that a cow belongs to a herd.

Ontology usage and evaluation

Query-based evaluation

As a preliminary evaluation of the DCPO, we adopted a query-based approach. The motivation behind was two-fold: (1) assess the practical usability of the populated ontology and (2) ensure the correctness of applied data transformations. Led by domain experts, we implemented SPARQL queries that reflect the typical questions experts might ask, e.g., to estimate the impact of cow management w.r.t to genetic potential. An example query is to compute average values on day 305 estimates for milk, protein, and fat for cows, herds, and regions for both production metrics (i.e., milk, fat, and protein) and estimated genetic potential (i.e., estimated breeding values). By substracting — relative to average — values for production and genetics, rough estimates of the quality of management practices for cows and herds are computed.

Generalized graph pattern mining with a DO

Structural regularities, or patterns, in the data can provide useful insights as to the general trends it reflects: They may lead an expert to discover unknown phenomena or, more realistically, to confirm an already formulated hypothesis. Therefore, such regularities, are worth mining and presenting to experts for an in-depth examination.

The immediate benefit of using a DO as vocabulary for pattern graphs is to enable the shared structure in data graphs to be explicitly described at the conceptual level, even though it may manifest in diverging ways at the data level. In other words, isomorphic graphs on the data level with diverging vertex and edge labels, which are thus, seemingly, unrelated, can become identical once their respective labels are generalized to the respective classes and generic properties from the DO.

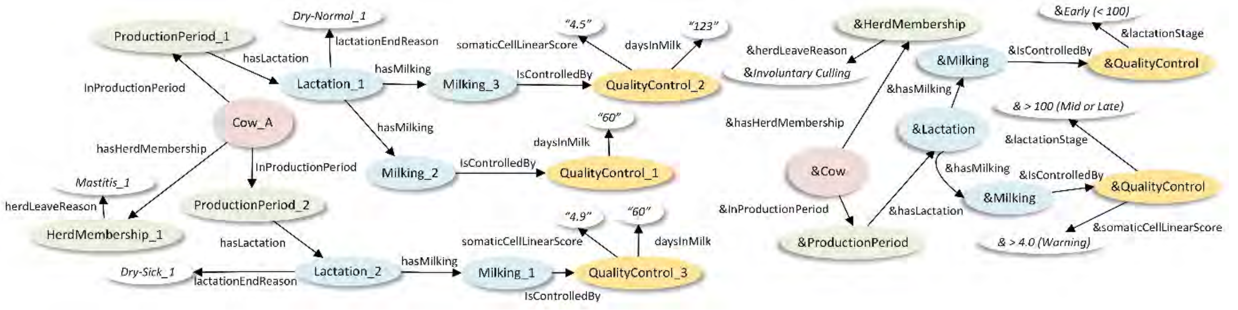


Figure 5. Example pattern (right) from a data graph (left), both supported by the dairy ontology.

Here, our DCPO and its instances act as a dual graph model where the former is used as a blueprint while the latter acts as the actual data to explore and analyze. Another way to picture it is to consider it as meta-data to formulate relevant hypothesis whereby graph data is used to (in-)validate such hypothesis.

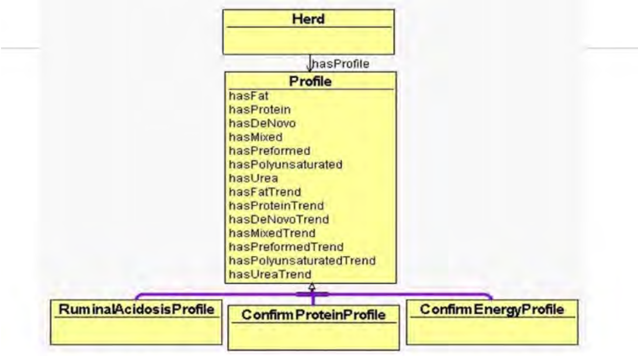
As an illustrative example, Figure 5 represents a data graph and a matching pattern that refers to DCPO. The pattern —found in an *ad-hoc* manner— was deemed potentially useful by our experts. It reflects the fact that a number of cows culled for reasons that were not under farmer's control (involuntary culling) had, prior to that event, at least one lactation with two quality controls, one of which indicates worrisome values of somatic cells. Such a co-occurrence is perfectly plausible as increased somatic cell counts are major signals for *mastitis* (inflammation of the udder tissue). Consequently, larger patterns contextualizing recurrent health issues could very well reveal the actual trigger for the involuntary culling. Therefore such patterns deserve to be investigated so that the underlying phenomena could be better understood and, if necessary, more closely monitored.

Experts' rules implementation

In the ontology, we describe what things exist in the domain, their attributes, and relationships. In general, this knowledge is external to animal experts, it is mostly established knowledge about the dairy process and its validity is not expected to change with the addition of new knowledge. But exists also internal knowledge developed by local experts. This internal knowledge usually expresses conditions on the state of individuals under which certain phenomena occurs. We use rules as knowledge representation in this case.

An example of this situation is the diagnosis of positive or negative effects of nutrition, management, and environmental factors in herds according to their bulk tank milk component profiles. The specification of the intervening entities (herd, milk profiles), their attributes (the values of the different measures for each profile component) and the relationships between them are expressed in the ontology (see Figure 6).

Animal experts have determined associations between extreme values of profile measures and different anomaly situations in herds. To express this associations, we define rules that use the vocabulary in the ontology. In this way, the ontology, rules, and a reasoner constitute a decision support system (DSS). Figure 6 shows an example of such rules: when a profile record is introduced in the ontology as an instance of a profile, the reasoner tries to match rule antecedents with instance data, if a match holds, the rule triggers and its consequent is executed, in this case it reclassifies the profile as a profile specialization.



```
hasFat=X and hasProtein=Y and... -> RuminalAcidosisProfile
hasUreaTrend=W and hasDeNovo=Z and... -> ConfirmEnergyProfile
...
```

Figure 6. Part of the ontology representing bulk tank milk profiles for herds.

Conclusion

We reflect here on our efforts on the design and implementation of DCPO, unifying several key aspects of dairy production. A major challenge we faced was the trade-off between plausible domain modeling and support for expressive knowledge discovery tools. At the current stage, it proved possible to reach both goals within a unique ontlogy. Next, we shall look at how to exploit ontology design patterns [18] and given we conform to the BFO, we envision an integration to the OBO (www.obofoundry.org) ontologies, with alignments to the relevant ontologies of the library. In longer run, we shall look at enhancing the data-centered ontology with knowledge discovered from the data by mining tools.

References

[1] **C.-J. Su and S.-F. Huang**, “Ontology-supported knowledge management with case-based reason-ing for intelligent health projection,” in BIBE, 2018, pp. 1–4.

[2] **C. G. Frasco and others**, “Towards an effective decision-making system based on cow profitability using deep learning,” in ICAART 2020, 2020, vol. 2, pp. 949–958.

[3] **J. P. C. Verhoosel and J. C. Spek**, “Semantics for big data applications in the smart dairy farming domain,” in Precision Dairy Farming Conference, Leeuwarden (NL), 2016.

[4] **D. Martini and others**, “agroRDF as a Semantic Overlay to agroXML: a General Model for Enhancing Interoperability in Agrifood Data Standards,” in CIGR Conference on Sustainable Agriculture through ICT Innovation, 2013.

[5] **S. D. K. Tomic and others**, “agriOpenLink: Towards Adaptive Agricultural Processes Enabled by Open Interfaces, Linked Data and Services,” in MTSR, 2013.

[6] **C. Caracciolo and others**, “The AGROVOC linked dataset,” Semant. Web, vol. 4, no. 3, pp. 341–348, 2013.

[7] **F. Kramer and T. Beissbarth**, “Working with ontologies,” in Bioinformatics, 2017, pp. 123–135.

[8] **C. C. Aggarwal and H. Wang**, Eds., Managing and Mining Graph Data, vol. 40. Springer, 2010.

[9] **C. C. Aggarwal and J. Han**, Frequent Pattern Mining, 2014th ed. Springer, 2014.

[10] R**. Srikant and R. Agrawal**, “Mining generalized association rules,” Futur. Gener. Comput. Syst., vol. 13, no. 2—3, pp. 161–180, 1997.

[11] **M. Adda and others**, “Toward {Recommendation} {Based} on {Ontology}-{Powered} {Web}-{Usage} {Mining},” Internet Comput. IEEE, vol. 11, no. 4, pp. 45–52, Aug. 2007.

[12] **T. Jiang and others**, “Mining generalized associations of semantic relations from textual web content,” IEEE TKDE, vol. 19, no. 2, pp. 164–179, 2007.

[13] A**. Cakmak and G. Ozsoyoglu**, “Taxonomy-superimposed graph mining,” in 11th EDBT, 2008, pp. 217–228.

[14] **U. Kursuncu and others**, “Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning,” in AAAI-MAKE Symposium, 2020.

[15] **H. Wang and others**, “Ontology-based deep restricted boltzmann machine,” in DExA, 2016, pp. 431–445.

[16] **T. Martin and others**, “Bridging the gap between an ontology and deep neural models by pattern mining,” in CSSA@CIKM, CEUR Workshop Proceedings, 2020, vol. 2708.

[17] **M. Rouane-Hacene and others**, “Relational concept analysis: mining concept lattices from multi-relational data,” Ann. Math. Artif. Intell., pp. 1–28, 2013.

[18] **A. Gangemi and V. Presutti**, “Ontology design patterns,” in Handbook on ontologies, Springer, 2009, pp. 221–243.