ATOL: The Multi-species Livestock Trait Ontology

Wiktoria Golik, Olivier Dameron, Jérôme Bugeon, Alice Fatet, Isabelle Hue, Catherine Hurtaud, Matthieu Matthieu.Reichstadt@inrae.Fr Reichstadt, Marie-Christine Meunier-Salaün, Jean Vernet, Léa Joret, et al.

ATOL: the multi-species livestock trait ontology

Wiktoria Golik[1], Olivier Dameron[2], Jérôme Bugeon[3], Alice Fatet[4], Isabelle Hue[1], Catherine Hurtaud[5], Matthieu Reichstadt[6], Marie-Christine Salaün[5], Jean Vernet[6], Léa Joret[3], Frédéric Papazian[1], Claire Nédellec[1], and Pierre-Yves Le Bail[3]

[1] INRA, UR1077 Jouy-en-Josas, France
[2] INSERM, UMR936 Université de Rennes1, France
[3] INRA, UR1037 Rennes, France
[4] INRA-CNRS, UMR0085 Université F. Rabelais, Tours, France
[5] INRA, UMR1348 Agrocampus Ouest, Rennes, France
[6] INRA, UMR1213 VetAgro Sup, Clermont-Ferrand, France

Abstract. This paper presents the multi-species Animal Trait Ontology for Livestock (ATOL) and the methodology used for its design. ATOL has been designed as a reference source for indexing phenotype databases and scientific papers. It covers five major topics related to animal productions: growth and meat quality, animal nutrition, milk production, reproduction and welfare. It is composed of species-independent concepts subsuming species-specific ones so that cross-species and species-specific reasoning can be performed consistently. In order to ensure a large consensus, three complementary approaches have successively been applied to its design: reuse of existing ontologies, integration of production-specific livestock traits by a large team of domain experts and curators and terminology analysis of scientific papers. It resulted in a detailed taxonomy of 1,654 traits that is available at `http://www.atol-ontology.com`

Keywords: animal trait, livestock, ontology, terminological analysis

1 Introduction

A phenotype is a set of values of the observable traits that characterize the animal at the molecular, physiological, anatomical, morphological or ethological levels. For example, an organism has the phenotype "blue" associated with the trait "eye color". Phenotypes are determined by multiple factors: simple genotypes determine eye color and complex genotypes interacting with environmental conditions determine size or behaviors. Observations and analysis of phenotypes are essential for both the understanding by physiologists of the conditions that produce phenotypes of interest and the selection effort conducted by geneticists. Animal selection has been performed empirically since domestication, and more rigorously since Mendel. It consists of improving a race by limiting the breeding to animals with the desired phenotypes. One of the major stakes for life sciences to become integrative and predictive is the ability to uniformly describe the traits (markers and effectors) that determine the phenotypes

of interest. The evolution of life sciences over the last two decades generated a deluge of data [**?**] that concerns many levels of biology that have potential implications for phenotypes [**?**] (particularly genomics, epigenomics, transcriptomics, proteomics and metabolomics). Bio-ontologies are an essential part of information systems because they support data integration and analysis across multiple levels of biology [**?**]. In this paper, we describe the ATOL multi-species livestock trait ontology, the motivation and its design method.

2 Background

In this section, we identify the main challenges to the integration of livestock production traits data and we survey previous efforts based on ontologies.

2.1 Data integration

Phenotype-related data is scattered across multiple databases, which makes their integration and their processing difficult. They are produced in numerous organizations, and each of these organizations is likely to harbor heterogeneous data structures. The databases typically have different models even when they refer to the same kind of information, different field names and different representations. For example, a first database can contain a column "weight" representing the weight in kilograms of a trout at three months, whereas a second database would contain a column "mass", a column "species" and a column "age" representing the weight in grams of animals from several species of different ages. Both databases fit the requirements of the daily internal activity of their producers, but their integration or their reuse in another context requires *ad-hoc* domain specific handling, independently of the unit conversion issues. The underexploitation of the phenotype data is the consequence of the lack of interoperability. It also hinders the progress of phenotype-related activities.

In practice, the conversion of all the existing databases into a unifying framework is impossible, assuming that such a framework would be unique. The classical solution for addressing heterogeneity consists of annotating data with metadata, i.e. describing them explicitly using a common formal framework [**?**]. In our previous example, this would mean that metadata indicate that the "weight" column from the first database and the "mass" column from the second database refer to the same entity and similarly it would normalize species and units. Metadata use offers a lightweight and flexible solution that does not require the modification of existing data to achieve at least partial interoperability. The first step consists of considering the existing databases for the definition of a common schema of metadata, then of defining an identifier for each notion of interest and finally of using these identifiers to describe the existing data. Each identifier can be associated to preferred terms and synonyms, possibly in multiple languages. This approach has been successfully used in the biomedical domain. In addition to a common framework for annotating data, it is also necessary to represent explicitly the generality relations between the annotations in order to reconcile and process automatically information with different levels of precision.

2.2 Ontologies

The explicit and formal description of livestock production traits and the relations between some of these traits constitute an ontology [**?**]. Several trait and phenotype ontologies are under active development. For phenotype measurement ontologies, see the review by Shimoyama et al. [**?**]. The Mammalian Phenotype Ontology (MPO) [**?**] is an OBO ontology describing phenotypes in a context of mutation and QTL studies in mammalian model species and human pathologies. It is mainly used for describing mouse and rat phenotypes. The Animal Trait Ontology (ATO) [**?**] is an ontology of traits for livestock and not of phenotypes. ATO provides a uniform vocabulary within one species as well as between species and it is used to annotate genomic data (for example QTL or SNP). The Vertebrate Trait Ontology (VT) [**?**] was created to provide a standardized vocabulary to facilitate the comparison of trait data within and across vertebrate species. It aims to describe vertebrate traits, defined as "measurable or observable characteristics", pertaining to the morphology, physiology, or development of an organism or its substructures. None of these ontologies fulfills the need for a reference source of metadata in the domain of multi-species livestock traits. ATO and VT partly covers the scope of ATOL. They are further detailed in sections 3.1 and 4.1.

2.3 Knowledge acquisition

The various methods applied to ontology modeling in specific domains mainly belong to three classes: reuse of existing ontologies, knowledge acquisition from experts and corpus-based acquisition [**?**]. The reuse ensures consistency and interoperability. Experts complement existing ontologies in order to fully cover the target scope. Document collections, i.e corpus, are also recognized as a rich source of knowledge as they provide terms that denote concepts, candidate to belong to the ontology. Corpus terms ensure a large coverage of the domain. They are also a source of alternative labels for naming the concepts. Term extractors automatically generate candidate terms when applied to a relevant set of documents. Among term extractors, BioYateA [**?**] is efficient [**?**] and well-adapted to the design of scientific ontologies. Despite the recent advances in term extraction and ontology learning, term candidates still need manual treatment. Termino-ontology editors support the construction of the ontology based on the terminology in a user-friendly way [**?**], [**?**]. TyDI fitted ATOL design needs because it supports expert collaborative work and direct expert interaction [**?**].

3 Methods

ATOL was developed in OWL format by a group of curators and domains experts using Protégé-4.1 and the WebProtégé collaborative environmentThe workgroup was composed of a leader, a biomedical ontology expert, five curators and about 50 domain experts. Each curator was in charge of one of the five topics according

to his/her domain of expertise. He/she managed a subgroup of domain experts. Special care was devoted to mix competencies in the subgroups and to balance expertise according to their scientific interest, fields and livestock species. INRA experts were motivated by the normalization effort to overcome their various laboratories and experimental farms specificities. Moreover, comparative physiology researchers conduct several collaborative programs on different species that needed to uniform definitions of phenotypic traits. We followed a three-step approach, (1) the reuse of ATO and VT, (2) the extension with livestock production specific traits (Section 3.1) and (3) the revision based on Animal Journal analysis (Section 3.2 and 3.3). Each step was done in close collaboration with James Reecy's group from Iowa University in order to maintain compatibility with the two ontologies, ATO and VT.

3.1 Construction of the initial version by the curators

First, each curator performed an extraction of the potentially relevant subtrees of the March 6, 2009 version of ATO and VT. Then, the curators and their respective experts subgroups selected the relevant concepts in the extraction and reviewed their definitions. This review phase was carried out in coordination with the ATO and VT team. For the sake of interoperability, references to the original concepts were preserved. Therefore, ATOL is aligned with ATO and VT by construction. Finally, the curators and their experts subgroups enriched the ontology by adding new concepts and by organizing them in a sound taxonomy. ATOL is composed of species-independent concepts subsuming species-specific ones. Each expert subgroup determined which species each concept could be associated with.

3.2 Analysis of corpus coverage by ATOL

Ontology modeling based on expertise has been usefully complemented by the study of a corpus of scientific international papers published in the animal trait domain conducted by a terminologist. The motivation was first to validate the terms chosen by experts as concept labels by checking their use in the literature. We chose the *Animal journal* because its scope includes all ATOL topics and beyond. We used the v1.0 early version of ATOL (April 2010) in order to evaluate the benefit of the corpus-based approach for the design of the next versions. This version contained 1,373 labels. The Animal corpus consists of 697 papers. The mapping of the concepts to the corpus was done by a straightforward projection of the concept labels to the corpus strings. 570 (42%) ATOL labels were found in the corpus. The high percentage of the matched terms was unfortunately due to many short and ambiguous labels that were too general (e.g. "performance", "approach") or incomplete (e.g. "pH"). They had to be rewritten and specialized accordingly. For instance, "pH" as descendant of "meat quality" should become "meat pH". Conversely, the terminologist identifies syntactic flaws that were easy to correct without involving deep expertise, including typographic errors,

translation errors, unnecessary conjunction of coordination (e.g. “and”) and frequent non-alphabetic characters that prevented the occurrence of ATOL labels in the corpus that were unnecessary in most of the case. This first analysis of ATOL labels led to a systematic correction reported in (Section 4.2).

3.3 Linguistic approach

A deeper linguistic terminological analysis was needed for suggesting further revisions of the ATOL labels that were not found in the corpus. It compared ATOL concept labels to the terms extracted from Animal journal papers.

Improvement of concept labels by linguistic variation. Among the 2,550 labels and synonyms found in ATOL version 3.5.8, only 922 occurred in scientific papers as measured by using Google Scholar hits. The manual examination of a subset of labels with 0 or rare occurrences showed that a major source of discrepancy was the choice of rare forms as concept labels over alternative names actually preferred by the authors of papers. Hopefuly many synonyms in the corpus were direct morpho-syntactic and semantic variations of the concept labels, such as “consumption of water” versus “water intake”. In this example, “water intake” is obtained by the permutation of the nouns of “consumption of water” and the replacement of “consumption” by its synonym “intake. We used FastR [**?**] for automatically computing such variations from ATOL labels with the goal of discovering relevant variants. Section 4.3 details the result of the application of FastR and its use for ATOL improvement.

Terminological analysis of the significance of ATOL labels. The variants of most of the long labels over 3 words were out of reach of FastR variations. We wanted then to discover new terms in the corpus that were synonym of the concept labels but not direct variations. We performed an extensive term extraction on the Animal corpus that provided many candidate terms for renaming these concepts among which the experts had to select the relevant ones.

We used BioYateA [**?**] for term extraction, after syntactic analysis by AlvisNLP [**?**]. BioYateA was provided with ATOL as source of certified terms. The extraction yielded 144,928 candidate terms. TyDI (Terminology Design Interface)) [**?**] assisted the manual exploration of the candidate terms and their matching to ATOL labels. For each label that was absent from the corpus, TyDI displayed the corpus terms that shared common features with the label and that could possibly be synonyms. The selection of the relevant features is done interactively. For instance, “withdrawal reflex” label had no match in the corpus and no FastR variant. The user enters queries such as “withdrawal” as term argument into TyDI interface. It displays 7 terms among which “withdrawal response” and “withdrawal reaction” are relevant related synonyms. The number of occurrences and the context help to select the most relevant and less ambiguous (see [**?**] for more details). The new term is then added as the preferred name for the concept in the ontology displayed by TyDI. The results of the use of BioYateA and TyDY for ATOL design is detailed in section 4.3.

ATOL extension by corpus-based term extraction. The lexical corpus-based approach supported by term extraction has also been applied to populate ATOL with new concepts. It differed from the previous case in that the experts used corpus term extraction from the beginning to design a whole ontology subtree, instead of using it for *a posteriori* revision. They looked for terms denoting new concepts on a given subject starting from representative words searched by TyDI. This work aimed at evaluating TyDI usability by a domain expert without the assistance of a knowledge engineer. Section 4.3 details the results.

4 Results

4.1 Initial version construction by the curators

The design of ATOL started with the reuse of existing ontologies, in particular ATO and VT. The 06 March 2009 version of ATO and VT was composed of 4,182 concepts, 3,692 (88 %) of which had a textual definition. Each curator and subgroup of experts selected the subtree of potential relevance for their domain of interest. Next, they manually enriched and organized their branch of interest. During these three steps, the ATO and VT parallel evolutions were monitored so that their changes could be propagated to ATOL. Conversely, the concepts added to ATOL by the experts were proposed for review to the ATO and VT experts. Figure 1 presents the composition and overlap of the five topics during the automatic extraction of concept from ATO and VT, the manual selection of the relevant ones and the addition of new concepts in version 4.4 of ATOL. The decreasing number of concepts of the growth and meat quality topics along the three steps can be explained by the fact that many concepts from ATO and VT were related to a specific muscle and sometimes to non edible muscle (eye muscle for example). They were first automatically extracted, but the focus of ATOL led us to manually exclude them. On the contrary, only few concepts related to milk production were present in the initial extraction and this topic was then notably extended in ATOL.

During the enrichment phase, a particular effort was devoted to organizing the ATOL ontology as a sound taxonomy, i.e. each class is formally a kind of its parents. For example, "adipose tissue fatty acid content" (`atol:0074`) and "adipose tissue lipid oxydation" (`atol:0075`) are two siblings subclasses of "adipose tissue lipid quality" (`atol:0073`). Thus, the superclass features logically hold for each of its subclasses and the subclasses of the subclasses by inference. Heritage allowed us to simplify modeling by factoring common features. It supports automatic reasoning so that if given data is annotated by a concept, one can infer that it is also annotated by all the ancestors of this concept since they are more general. This is used to reconcile data with different levels of precision. Whenever necessary, we also used multiple inheritance by assigning more than one superclass to a class. For example, "body weight" (`atol:0351`) is a subclass of both "animal performance trait" (`atol:1516`) and "growth trait" (`atol:0855`). Table 1 presents the distribution of the concepts among topics and their overlap.

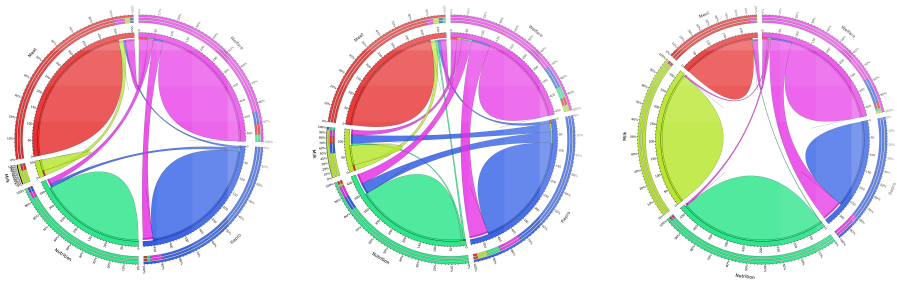

Fig. 1. Composition and overlap of the five branches of ATOL: meat growth and quality (red, top left), milk (light green, left), nutrition (dark green, bottom left), reproduction (blue, bottom right) and welfare (purple, top right) at the three stages of ATOL design, (1) extraction from ATO and VT, (2) manual selection of relevant concepts and (3) ATOL v4.4 after enrichment.

?	**Repro**	**Milk**	**Meat**	**Welfare**	**Nutrition**
Repro	274	0	0	67	0
Milk	0	420	0	5	0
Meat	0	0	228	15	2
Welfare	67	5	15	331	6
Nutrition	0	0	2	6	462

Table 1. Concept distribution by topic after manual enrichment of ATOL version 4.4.

The structuring phase resulted in concepts previously shared between welfare and nutrition being assigned to either of the two domains (6 shared concepts), whereas the concepts shared between welfare and reproduction (67 concepts) remained common. During the selection and the enrichment phases, the experts determined for each concept the list of species they were relevant for. Figure 2 shows the distribution of shared concepts between cow and sheep (left, 93 % of common traits) and cow and trout (right, 51 % of common traits) for each ATOL topic. Not suprisingly cows and sheep globally share the same traits. Cows and trout share the meat quality traits and are less similar otherwise. Obviously, milk-related traits are cow-related and have no counterparts in trout. This illustrates the genericity of ATOL traits among species.

4.2 Analysis of corpus coverage by ATOL

The extensive shallow analysis of ATOL labels with respect to the Animal Journal yielded 156 new concepts or synonyms in ATOL 1.0 (10% increase) and among them, 27 were present in the corpus. This work led to clear guidelines about the form of the labels that curators should apply to the future versions of ATOL. We then measured the improvement of label quality in version 3.5.8 of ATOL (Dec. 2011) that followed the guidelines and included many new traits. Only 2% of the 2,550 labels had typographic errors. The measure of their occurrence in the literature reached 43%, a much higher rate than previously, which demonstrates the benefit of a corpus-based evaluation of the ontology.

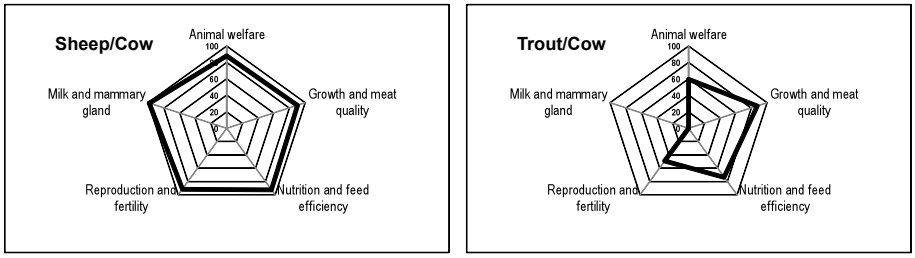

Fig. 2. Number of common traits between cow and sheep (left) and cow and trout (right) traits for the five subtree of ATOL.

4.3 Linguistic approach

Improvement of concept labels by linguistic variation. Compared to straightforward coverage analysis, the application of FastR led to many revisions based on morpho-syntactic analysis. All 1,605 ATOL labels without any occurrence in the corpus were given to FastR together with the corpus of 697 papers and WordNet as a source of semantic variations [**?**]. Table 2 gives the most frequent variants with their frequency. It is noticeable that in many cases, the variant was the most frequent form but not necessarily the less ambiguous as "slaughter age" instead of "age at slaughter". The table illustrates the diversity of the lexical relations between the labels and their variants. They are not all synonyms but also hyper- or hyponyms that may be relevant to ATOL.

ATOL label	**#occ**	**Corpus variant**	**#occ**
milk yield	1192	milk production	1485
energy expenditure	48	energy intake	291
meat trait	10	meat quality trait	194
age at slaughter	52	slaughter age	133
parental behaviour	0	maternal behaviour	104
milk yield	1192	milk fat yield	85
water intake	147	water consumption	76
feeding behaviour	0	feeding behavior	71

Table 2. ATOL original terms and most frequent variants proposed by FastR.

FastR computed the label variants from corpus terms by applying variation rules that performed insertion, permutation and replacement of words by WordNet synset members. It yielded 1,190 pairs of ATOL labels – variants for 218 different labels. Among them a knowledge engineer validated 541 synonymy pairs for 171 different labels, excluding specializations and other relevant terms that were not strict synonyms. Semantic variation that is due to WordNet is involved in 60% of the positive pairs (50% of the total) demonstrating the clear benefit of semantic variation and the use of external resource. Among the 1,605 labels without any occurrence in the corpus, FastR then automatically found relevant alternative names for 10% (171) of them. The computation of term variants by

linguistic analysis appeared as a valuable solution for improving concept names especially short names. Their inclusion in ATOL is in progress.

Terminological analysis of the significance of ATOL labels. After synonym computation, 1,434 ATOL labels still remained with 0 occurrences in the corpus. We used BioYateA and TyDI to analyze the reasons why so many long ATOL labels are absent from the corpus. The lessons from this first terminological study are various. (1) The paper corpus should be extended to journals other than Animal in order to explore a larger set of candidate terms. The *Journal of Animal Science* and *Livestock Science* are obvious candidates relevant to the scope of ATOL. (2) Some synsets are missing in WordNet that are very relevant to the Animal domain. Providing FastR with them would enable it to compute many additional relevant synonym variants. Among the most frequent related synonyms, "content"/"concentration" occurs in 857 of the 0–occurrence labels and "meat"/"flesh" in 31. This would enable to compute for instance "adipose tissue vitamin content"$_L$ / "vitamin concentrations in adipose tissue"$_T$ or "flesh physicochemical trait"$_L$ / "physicochemical properties of meat"T_T . Such frequent synonyms in ATOL should be considered in order to improve WordNet power. (3) The animal product is always mentioned in the trait label, e.g. "Meat" and "Milk" frequently occurred in 263 of the missing labels. Automatically removing the product name from the labels yielded many hits in TyDI; thus proving the relevance of those labels although the matched terms were not synonym. For instance, "milk color redness" is not synonym of "meat color redness", but the presence of "color redness" in the text is a good indicator of the use of the redness concept. (4) Animal names are frequently inserted in corpus terms, as in "average daily gain"$_L$ / "average pigs daily gains"$_T$ preventing the label from being found. However, the occurrence of such more specific terms confirms the relevance of the label. The matching process can be automated, first by using the list of animals associated to the concepts in the ontology in the form of subsets, then by designing an extensive list of their variant names. The remaining cases are due to paraphrases: the concept is not expressed by a term but by a more complex construction. Corpus term analysis combined with the semantic search engine AlvisIR [**?**] helps in finding these paraphrases but their association to ATOL labels cannot be fully automated.
Example. "Seasonality of female sexual activity"$_L$ / "Decreasing photoperiod plays an important role in activating sexual activity in seasonal breeders"$_T$.
The terminological analysis of terms that are close to ATOL labels yielded promising new directions for automatically identifying ATOL concepts in the corpus. Their relevance will be assessed in the future by quantitative measures of ATOL label occurrences in a larger corpus.

ATOL extension by corpus-based term extraction We experimented with the corpus-based approach described in section 3.3 for creating new concepts in the feed domain. An expert was taught how to use TyDI. He measured the relevance of the terms proposed by TyDI, according to their frequency and de-

cided accordingly wether to create the corresponding concept or not. The words "nutrition", "feed" and "flux" that are representative of the topic were first searched through TyDI interface. It yielded 847 terms among which a subset has been used to design the nutrition subtree. Then synonyms of these concepts have been searched and added to ATOL. TyDI was then particularly useful to evaluate which of the forms is the most popular, e.g. "nitrogen content in feed" versus "nitrogen content of feed". It was then used for enriching the ontology by systematically looking for all specific arguments of a given concept. For instance, digestibility is a main concepts in nutrition. TyDI supported the search for all nutriments to which digestibility applies, (e.g. nitrogen, phosphorus, fiber). It yielded about thirty words. The search for the organs where the digestibility is measured (e.g. rumen, intestinal tract, cloacae) yielded 16 new concepts. This experiment confirmed that TyDI tool as a valuable solution for supporting corpus-based terminological analysis for ontology design.

5 Discussion

The current version 4.6.8 of ATOL defines 1,656 concepts among which 1,186 are specific to ATOL. 545 concepts are shared with VT and 341 VT concepts were annotated by the ATOL group. ATOL fills a gap in the domain of trait and phenotype ontologies such as ATO and VT that have different scopes. Their structure was not compatible with ATOL requirements preventing extension to livestock. ATO recently evolved towards a consortium of ontologies on products (PT), Animal breed ontology on species, and VT. Originally, VT was intended to describe model species traits like those of mice and rats. Its organization follows an academic point of view (e.g. morphology, functions) without reference to species. Its further extension to livestock species via ATO retained this hierarchical perspective. Moreover, VT only considers directly measurable traits (called simple traits). It excludes complex traits that are defined from other simple or complex traits such as gonado-somatic ratio or body mass index. ATOL focuses on the different kinds of animal products (quantity and quality of meat, milk and eggs) or of breeding (alimentary efficiency, fertility, welfare). These domains rely on numerous complex traits used by both breeding professionals and researchers. However, the VT, ATO and ATOL leaders agreed to shared as many traits as possible using explicit cross-references. This solution both preserves the specific traits and organization of ontologies, and maximizes interoperability. In the current version of ATOL, the traits are organized in a is-a hierarchy. We plan to include additional relations such as part_of to represent composition, as well as is_an_indicator_of and is_a_standardization-of to take into account the connection between a trait of interest and the different modes of observation of this trait. Human effort made by INRA for the specific development of ATOL and its sustainability over time is part of its strategy to develop operational integrative and predictive biology approaches for the systemic management of livestock in France and Europe. Since the project start in 2009, ATOL development is estimated at 62 man-months. The second phase of maintenance and evolution

of ATOL is estimated at 3 man-months that will be spread over a group of 10 persons from INRA. Among the programs in which ATOL is used, AQUAEXCEL [18] is an example in which the fish-related part of ATOL is reused for resource sharing and normalization among partners, notably for fish models and experimental methods. Conversely, improvements suggested by AQUAEXCEL are propagated into ATOL. User feedback through the ATOL website is welcome. As ATOL gains acceptance, it will be important to follow international standard for ontology design and description

The design of ATOL has shown that the terminological analysis was more efficient when used during the design of the ontology as done for nutrition, than *a posteriori*. This is the consequence of both methodological reasons and expert motivation. When the experts considered the design achieved, the terminological analysis appeared more as a corrector that revealed flaws than as a useful support for finding new concepts or the best way to express them. For the development of the new parts of ATOL, such as environmental factors, the terminology-based approach will be used from the very beginning and fully integrated into the methodology. The addition of synonyms to ATOL from the corpus opened new perspectives: it made ATOL usable for full-text indexing of the Animal journal by the semantic search engine AlvisIR. A preliminary public version is available at [**?**]. Semantic search fully takes advantage of the hierarchical structure of ATOL. For instance, the query "milk composition trait" retrieves 77 articles that mention specific traits such as "milk fat concentration". The query on Google Scholar does not retrieve any answer. The Google Scholar query "milk composition" without "trait" retrieves only 24 papers from the same collection. The query "meat quality" yields 318 hits in AlvisIR, 71 in Google Scholar. These two examples illustrate the added value of ATOL for semantic search. In the near future, the extension of ATOL by the terminology level will be achieved, thus making the *Animal search engine* fully operational.

6 Conclusion

This paper has presented ATOL, a multi-species livestock trait ontology. It has been designed as a reference source for phenotype databases and scientific papers metadata. ATOL covers five major topics related to animal product: growth and meat quality, animal nutrition, milk production, reproduction and welfare. The initial design phase relied on groups of experts and curators. This ensured a general coverage of each five topics and that concepts were organized in a sound taxonomy. A terminological analysis of the Animal Journal was then conducted in order to identify and rename irrelevant concept labels and to identify new concepts. It improved ATOL at different levels of conceptualization. In addition the terminological analysis validated the relevance of ATOL as a resource for the automatic semantic indexing of literature.

References

1. JA. Blake and CJ. Bult. Beyond the data deluge: Data integration and bio-ontologies. *Journal of Biomedical Informatics*, 39(3):314–320, 2006.
2. J-F Hocquette, C. Capel, V. David, D. Guéméné, J. Bidanel, C. Ponsart, P-L Gastinel, P-Y Le Bail, P. Monget, P. Mormède, M. Barbezant, F. Guillou, and J-L Peyraud. Objectives and applications of phenotyping network setup for livestock. *Animal Science Journal*, 83:517–528, 2012.
3. J. J. Cimino and X. Zhu. The practical impact of ontologies on biomedical informatics. *Methods of information in medicine*, 2006.
4. NH Shah, C Jonquet, AP Chiang, AJ Butte, R. Chen, and MA Musen. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC bioinformatics*, 10 Suppl 2:S1, 2009.
5. JBL Bard and SY Rhee. Ontologies in biology: design, applications and future challenges. *Nature reviews. Genetics*, 5(3):213–222, 2004.
6. M. Shimoyama, R. Nigam, L. Sanders McIntosh, R. Nagarajan, T. Rice, DC Rao, and MR Dwinell. Three ontologies to define phenotype measurement data. *Frontiers in genetics*, 3:87, 2012.
7. CL Smith, C-AW Goldsmith, and JT Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7, 2004.
8. L M Hughes, J Bao, Z-L Hu, V Honavar, and J M Reecy. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *Journal of animal science*, 86(6):1485–1491, 2008.
9. C. Park, SM. Bello, C. Smith, Z-L Hu, D. Munzenmaier, M. Shimoyama, J. Eppig, and J. Reecy. The vertebrate trait ontology: A controlled vocabulary to facilitate cross-species comparison of trait data. In *Proceedings of the Plant and Animal Genomes, 20th Conference. San Diego, CA*, 2012.
10. M. Uschold and M. King. Towards a methodology for building ontologies. In *IJCAI-95 workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.
11. C. Nédellec, W. Golik, S. Aubin, and R. Bossy. Building large lexicalized ontologies from text: a use case in indexing biotechnology patents. In *Proceedings of EKAW, Portugal.*, 2010.
12. T. Mondary, A. Nazarenko, H. Zargayouna, and S. Barreaux. The quaero evaluation campaign on term extraction. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*, 2012.
13. N. Aussenac-Gilles, S. Després, and S. Szulman. The terminae method and platform for ontology engineering from texts. In *Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text*, pages 199–223, 2008.
14. C. Jacquemin. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, chapter A symbolic and surgical acquisition of terms through variation, pages 425–438. Springer, Heidelberg, 1996.
15. C Nédellec, A Nazarenko, and R Bossy. *Ontology Handbook*, chapter Information Extraction, pages 663–686. Springer-Verlag, 2008.
16. G. A. Miller, R. Beckwith, C. D. Fellbaum, D. Gross, and K. Miller. WordNet: An online lexical database. *Int. J. Lexicograph.*, 3(4):235–244, 1990.
17. R. Bossy, A Kotoujansky, W. Golik, S. Aubin, and C. Nédellec. Close integration of ML and NLP tools in BioAlvis for semantic search in bacteriology. In *Proc. of the Workshop on Semantic Web Applications and Tools for Life Sciences*, 2008.
18. AlvisIR for ANIMAL. http://bibliome.jouy.inra.fr/test/alvisir/animal/.