INRA Prod. Anim.,
2014, 27 (3), 195-208

Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL

P.-Y. LE BAIL[1], J. BUGEON[1], O. DAMERON[2], A. FATET[3,4,5,6], W. GOLIK[7], J.-F. HOCQUETTE[8,9], C. HURTAUD[10, 11], I. HUE[12], C. JONDREVILLE[13], L. JORET[1], M.-C. MEUNIER-SALAÜN[9], J. VERNET[8,9], C. NEDELLEC[4], M. REICHSTADT[8,9], P. CHEMINEAU[3,4,5,6]

[1] INRA, UR1037 LPGP, F-35000 Rennes, France
[2] Université de Rennes 1, IRISA, Dyliss team, F-35000 Rennes, France
[3] INRA, UMR85 PRC, F-37380 Nouzilly, France
[4] CNRS, UMR7247, F-37380 Nouzilly, France
[5] Université François Rabelais de Tours, F-37000 Tours, France
[6] IFCE, F-37380 Nouzilly, France
[7] INRA, UR1077 MIG, F-78352 Jouy-en-Josas, France
[8] INRA, UMR1213 Herbivores, F-63122 Saint-Genès-Champanelle, France
[9] Clermont Université, VetAgro Sup, UMR 1213 Herbivores, BP 10448, F-63000 Clermont-Ferrand, France
[10] INRA, Agrocampus Ouest, UMR1348 PEGASE, F-35590 Saint-Gilles, France
[11] Agrocampus Ouest, UMR1348 PEGASE, F-35000 Rennes, France
[12] INRA, UMR1198 BDR, F-78352 Jouy-en-Josas, France
[13] INRA, Université de Lorraine, USC340 AFPA, F-54500 Vandœuvre-lès-Nancy, France
Courriel : pierre-yves.lebail@rennes.inra.fr

Depuis l'avènement des technologies à haut débit, la quantité d'informations disponibles sur les génomes et les performances zootechniques des animaux d'élevage, *via* les bases de données et les articles scientifiques, a explosé. Sans un langage de référence pour décrire ces productions, l'accès pertinent aux informations publiées et le partage effectif des données resteront partiels et imprécis, rendant impossible leur réutilisation pour d'autres analyses, y compris les plus globalisantes qui visent à l'amélioration des performances animales par la sélection génomique et les conduites d'élevage de demain.

Les sauts technologiques qu'a connus la biologie lors des deux dernières décennies ont produit des quantités inédites de données. Alors que le séquençage du génome entier de l'Homme a demandé plusieurs années et mobilisé de nombreux laboratoires avec un coût extrêmement élevé (plusieurs millions d'euros), aujourd'hui le même travail de séquençage hors assemblage ne demanderait que quelques jours à une équipe et à un coût modeste (bientôt de l'ordre du millier d'euros). Concernant les animaux d'élevage, ces avancées technologiques accroissent l'information génétique et génomique disponible et favorisent l'établissement de liens entre polymorphismes génétiques et variabilité phénotypique. Elles permettent également l'accès à d'autres niveaux de fonctionnement du vivant, en particulier le niveau moléculaire (génome, épigénome, transcriptome, protéome, métabolome…) dont les liens avec les phénotypes (encadré 1) apparaissent aujourd'hui plus étroits (Hocquette *et al* 2009, Monget et Le Bail 2009, Groth *et al* 2011).

Paradoxalement, alors que les phénotypes sont utilisés depuis des siècles, en particulier à des fins de sélection génétique (de Rochambeau 2007), les référentiels définissant les caractères phénotypiques des animaux d'élevage sont à la fois multiples et fragmentaires, malgré l'intérêt de la communauté internationale, représentée notamment par le groupe ICAR (« *International Committee for Animal Recording* »[1]) qui produit régulièrement un document intitulé « *International agreement of recording practices* » dans le cas des ruminants. Décrire de façon homogène, et si possible unique, l'ensemble des caractères, ou traits, phénotypiques (encadré 1) sur lesquels reposent les phénotypes d'intérêt (en les mettant en relation avec des particularités génétiques dans une perspective intégrative et prédictive), est l'un des enjeux actuels des sciences du vivant. Cet objectif nécessite que les caractères phénotypiques soient clairement définis, normalisés, mesurés et référencés avec précision (Hocquette *et al* 2012). Parmi les outils de standardisation à notre disposition, les ontologies apparaissent pertinentes car elles permettent d'intégrer des données de provenance ou de nature hétérogènes. Par exemple, « *Gene Ontology* »[2] permet de décrire de manière uniforme les produits des gènes de différentes espèces et donc de faire des comparaisons entre gènes, voire entre

[1] http://www.icar.org/
[2] http://www.geneontology.org/

Encadré 1. *Phénome, phénotype, caractère phénotypique.*

Le « phénome » d'un individu est l'ensemble des « caractères phénotypiques » (ou traits phénotypiques) observables (aux niveaux éthologique, morphologique, anatomique, physiologique, moléculaire) qui le caractérisent (couleur des yeux, des cheveux, taille maximale...). Par exemple, au caractère phénotypique « couleur des yeux » peut correspondre l'état de « phénotype » « bleu ». Dans ce cas, le phénotype est régi par un déterminisme génétique simple. La couleur des yeux est un caractère phénotypique discret (ou discontinu) puisqu'il présente un nombre de phénotypes limité (bleu, marron, vert…), contrairement à la taille d'un individu (caractère continu) qui varie de manière continue au sein d'un intervalle compris entre une valeur minimale et une valeur maximale. De manière générale, un même phénotype peut dépendre de l'expression de plusieurs gènes et de leurs interactions. Il est donc la résultante de l'expression du génotype modulée ou non par le milieu environnant. C'est par exemple le cas de la croissance d'un animal qui dépend de paramètres physiologiques intrinsèques (systèmes hormonaux, capacité digestive…), mais aussi des conditions d'élevage et du milieu ambiant (alimentation, exercice, température, interactions entre individus, niveau de stress, pathologie…). L'observation directe et simple des phénotypes a favorisé leur large utilisation par les physiologistes dans l'étude des fonctions biologiques et par les généticiens dans leurs programmes de sélection, par exemple avec le caractère « culard » de certains bovins qui correspond à une hypertrophie musculaire. Mais les phénotypes peuvent porter sur des caractères moins « visibles » comme la structure du muscle (nombre de fibres musculaires, surface de tissus adipeux…) afin d'améliorer la qualité de la viande (Hocquette *et al* 2006) et nécessiter une investigation plus invasive comme la biopsie ou la dissection. Enfin, pour rendre certains phénotypes plus pertinents, il est utile de les standardiser (caractère phénotypique dit « complexe » ou « calculé ») en les exprimant comme une fonction mathématique plus ou moins complexe de caractères simples (par exemple, le poids d'un organe peut être exprimé en % du poids du corps). Aujourd'hui, un caractère phénotypique peut être le caractère d'intérêt lui-même, un bio-marqueur (ou indicateur) qui lui est corrélé, ou un effecteur qui le contrôle.

espèces. De plus, la structure de « *Gene Ontology* » rend possible la caractérisation et l'optimisation de listes de gènes en faisant ressortir les fonctions partagées par ces gènes.

Concernant l'« *Animal Trait Ontology for Livestock* » (ATOL) développée à l'Inra, l'objectif est ici de présenter l'élaboration, la structure et les performances de cette ontologie dédiée aux caractères phénotypiques des animaux d'élevage.

1 / Un référentiel des caractères phénotypiques : pourquoi et pour qui ?

1.1 / Un référentiel pour mieux partager

a) Rendre compatible les bases de données traitant de concepts similaires

La mise en relation des bases de données (génétiques, phénotypiques…) pour l'ensemble du vivant, grâce à un seul et même outil, est aujourd'hui hors de portée en raison de la multiplicité des espèces, des niveaux biologiques étudiés, des évolutions des phénotypes dans le temps et l'espace, mais aussi des limites de nos connaissances. Par réalisme, nous devons limiter nos ambitions intégratives aux enjeux de chaque domaine d'application, un outil plus ciblé étant plus efficace et plus facile à maintenir dans le temps. Pour les filières animales (mammifères, volailles, poissons), il s'agira de cerner le périmètre des caractères permettant de maîtriser les productions (viande, lait, œuf…), la qualité des produits (sanitaire, nutritionnelle, sensorielle) et de respecter les attentes sociétales dans le domaine de l'environnement (rejets azotés, gaz à effet de serre…) ou du bien-être (faible perturbation de l'animal dans son environnement d'élevage).

De plus, les données relatives aux caractères phénotypiques de production animale sont hétérogènes, multiples, voire redondantes. Il existe ainsi dans différents centres de recherches, voire au sein d'un même centre de recherches, de nombreuses bases comportant des informations similaires, mais partageant rarement la même structure, proposant rarement des passerelles vers des bases semblables ou vers d'autres types de bases recensant des données acquises à un autre niveau biologique. Il arrive aussi souvent qu'un même nom de caractère ne recouvre pas exactement la même définition selon les auteurs, les espèces ou les protocoles de mesure. Par exemple, si deux bases de données ont comme champs respectifs « masse » et « poids » pour caractériser le poids d'un animal, il est impossible de déterminer automatiquement si ces données sont équivalentes, ont la même unité et peuvent être comparables. Si chacune de ces bases de données convient bien à une utilisation locale spécifique, la dénomination distincte des champs « poids » et « masse » limite la possibilité de les intégrer dans des bases plus étendues pour les partager, les comparer, les combiner ou les utiliser dans un autre contexte.

Pour résoudre ces problèmes de compatibilité entre bases, la solution classique consiste à annoter les données en utilisant un référentiel commun sous la forme de métadonnées (Corcho 2006). Dans l'exemple précédent, cela reviendrait à identifier que le champ « poids » de la première table représente le même concept que le champ « masse » de la seconde. Idéalement, un seul et même terme standard pourrait ainsi être universellement référencé. Dans la pratique, le plus souple et le plus efficace est de s'accorder sur un terme préféré (par exemple « masse ») et d'annoter toutes les données des deux champs avec ce même terme sans remettre en cause la structure des bases. Une telle approche, déjà utilisée dans le domaine biomédical (Bodenreider et Stevens 2006, Cimino et Zhu 2006) ou dans celui de la génomique (Blake et Bult 2006), a été tentée dans le domaine de la qualité de la viande bovine. Cependant, elle s'est avérée difficile à mettre en pratique pour la réalisation de méta-analyses[3] (Hocquette *et al* 2011). C'est pourquoi d'autres communautés (comme le « *Beef Cooperative Research Center* » en Australie) ont préféré définir dès le départ des méthodes standards d'appréciation de la qualité de la viande (Watson *et al* 2008), plus pertinentes sur la durée que l'utilisation des métadonnées.

b) Un référentiel commun organisé hiérarchiquement : un plus

Au regard des enjeux précédents, il devient important de pouvoir interroger et comparer des observations issues de diverses expériences, éventuellement stockées dans des bases de données différentes. Dans une certaine mesure, cela peut concerner aussi le partage d'informations relatives à un caractère phénotypique particulier, commun à différentes espèces animales. En plus de l'analyse automatique et facilitée de données d'origines variées, un tel système a intrinsèquement une valeur pédagogique, permettant à l'utilisateur de consulter les caractères phénotypiques spécifiquement liés à telle production animale, voire de comprendre leur élaboration.

Pour être robuste, le système envisagé doit à la fois *i)* fournir un référentiel commun à plusieurs sources de données, quelle que soit leur provenance et *ii)* permettre le traitement de données dont

[3] Une méta-analyse est une analyse statistique combinant les résultats de plusieurs études indépendantes autour d'un problème donné.

les degrés de précision diffèrent. Dans le premier cas, il s'agit de désigner les caractères phénotypiques de façon normalisée à l'aide d'un vocabulaire contrôlé ; dans le second, le référentiel - structuré par des relations entre certains caractères - permet d'automatiser la prise en compte de niveaux de granularité différents et les raisonnements qui en découlent. Par exemple, cela permet de déterminer qu'une base contenant des mesures relatives au caractère « contenu en acides gras des tissus adipeux » est pertinente pour une requête plus globale concernant l'ensemble des caractères relatifs aux « contenu en lipides des tissus adipeux ».

1.2 / Les ontologies comme référentiels

a) Comment est structurée une ontologie et selon quel format ?

Bard et Rhee (2004) définissent les ontologies comme une « façon formelle de représenter des connaissances en décrivant les concepts à la fois par leur sens et les relations qui les lient ». Les connaissances décrites, de manière précise et univoque, et formalisées dans une ontologie permettent ainsi l'interprétation automatique de données d'intérêt en exploitant des relations qui autrement ne seraient qu'implicites. En pratique, les ontologies sont principalement constituées de classes (ou « concepts » ou « types »), de relations (ou propriétés), éventuellement de règles de raisonnement. L'utilisation des classes et des propriétés pour décrire les données se fait *via* leur identifiant.

Par exemple, l'identifiant ATOL: 0000350 correspond au concept « *body height* », et ATOL:0000093 correspond à « *body weight* ». Comme tous deux sont des sous-classes d'ATOL:0000855 (« *growth trait* »), cela permet à un système exploitant les ontologies de traiter des données avec différents niveaux de précision. Enfin, puisque chaque identifiant spécifie l'ontologie à laquelle appartient le concept (ici « ATOL: »), cette structuration rend possible la combinaison de plusieurs ontologies pour décrire les données plus finement et sans ambiguïté.

Il existe deux principaux formats de représentation des ontologies :

- OBO (« *Open Biomedical Ontology* ») a été créée pour représenter « *Gene Ontology* » avant sa réutilisation pour d'autres ontologies ;
- OWL (« *Web Ontology Language* »), issu du Web Sémantique, est plus puissant et expressif qu'OBO mais plus difficile à mettre en œuvre.

La distinction entre les deux formats n'est pas toujours claire car ils possèdent une base commune. La plupart des ontologies au format OWL n'exploitent pas pleinement ses spécificités et peuvent donc être converties sans perte d'information au format OBO. Enfin, des initiatives comme « OBOFoundry » et « BioPortal » constituent des entrepôts permettant de centraliser les ontologies, « BioPortal »[4] permettant en plus de les interroger de manière globale.

b) Les ontologies existantes pour phénotyper le monde animal

Parmi les nombreuses ontologies existantes, seules quelques ontologies de phénotypes animaux étaient disponibles ou en cours de développement quand le projet ATOL a été initié :

- « *Mammalian Phenotype Ontology* » (MPO)[5] décrit les phénotypes de plusieurs mammifères dans le contexte d'études de mutations et/ou de QTL (Locus de caractères quantitatifs) impliqués dans des modèles biologiques (souris, rat) et/ou des pathologies humaines (Smith *et al* 2005) ;
- L'ontologie « *Human Phenotype Ontology* » (HPO)[6] ne décrit, quant à elle, que des phénotypes de pathologies humaines (Robinson et Mundlos 2010). Elle sert à annoter des données pour les réutiliser dans un contexte de diagnostic clinique ou d'analyse d'expression génétique associée à des maladies humaines ;
- L'ontologie « *Phenotypic quality ontology* » (PATO)[7] est une ontologie des qualificatifs des phénotypes (par exemple « rouge » pour décrire la couleur des yeux). Il ne s'agit donc pas d'une ontologie des caractères phénotypiques.

Il existait aussi des ontologies spécifiques décrivant l'anatomie d'espèces comme « *Zebrafish Anatomical Ontology* » (ZAO)[8] ou à dimension plus large comme « *Vertebrate Skeletal Anatomy Ontology* » (VSAO)[9].

Malgré leur intérêt pour la recherche animale, aucune de ces ontologies opérationnelles ne répondait aux attentes des acteurs des productions animales (Mungall *et al* 2010), mises à part quelques bases de caractères documentant une seule fonction, souvent limitées à un modèle animal, comme STORE-FISH pour la reproduction des poissons (Teletchea *et al* 2007) ou PigQTLdb pour la génétique porcine, où seuls les caractères phénotypiques en relation avec un QTL étaient répertoriés (Hu *et al* 2005).

1.3 / Quels besoins pour les productions animales ?

La direction du département PHASE (Physiologie Animale et Systèmes d'Elevage) de l'Inra s'était fixée comme objectif de développer un référentiel des caractères phénotypiques dédié aux animaux de rente pour encourager une recherche intégrative sur l'animal au sein de l'Inra. Ce référentiel devait répondre à un cahier des charges visant à favoriser *i)* l'établissement de relations précises entre phénotype et génotype, comme par exemple associer un phénotype à un allèle[10] ou une combinaison d'allèles ; *ii)* le repérage au cours de l'évolution de constantes fonctionnelles au sein de groupes de gènes afin d'en tirer des règles générales chez les vertébrés ; *iii)* le développement d'outils de modélisation et de prédiction de phénotypes, et des performances associées, en fonction des conduites d'élevage et *iv)* la construction de modèles *in silico* mimant la physiologie cellulaire et animale.

Pour l'Inra et ses partenaires professionnels, les phénotypes d'intérêt sont ceux qui permettent, à terme, de comprendre et d'orienter les pratiques d'élevage et la sélection animale dans un contexte de durabilité : produire mieux (en quantité et qualité) en respectant l'environnement et le bien-être animal. Ainsi, cette ontologie favorisera la mise en place d'un langage commun entre zootechniciens, physiologistes et généticiens, facilitera les projets collaboratifs entre disciplines et/ou modèles animaux différents, rendra les informations partageables par l'utilisation des caractères référencés dans les publications et les bases de données. Par ailleurs, cette ontologie pourra servir de support pédagogique dans le domaine des productions animales. Le référentiel recherché devra donc prendre la forme d'une ontologie des caractères phénotypiques, dans laquelle le choix des caractères sera

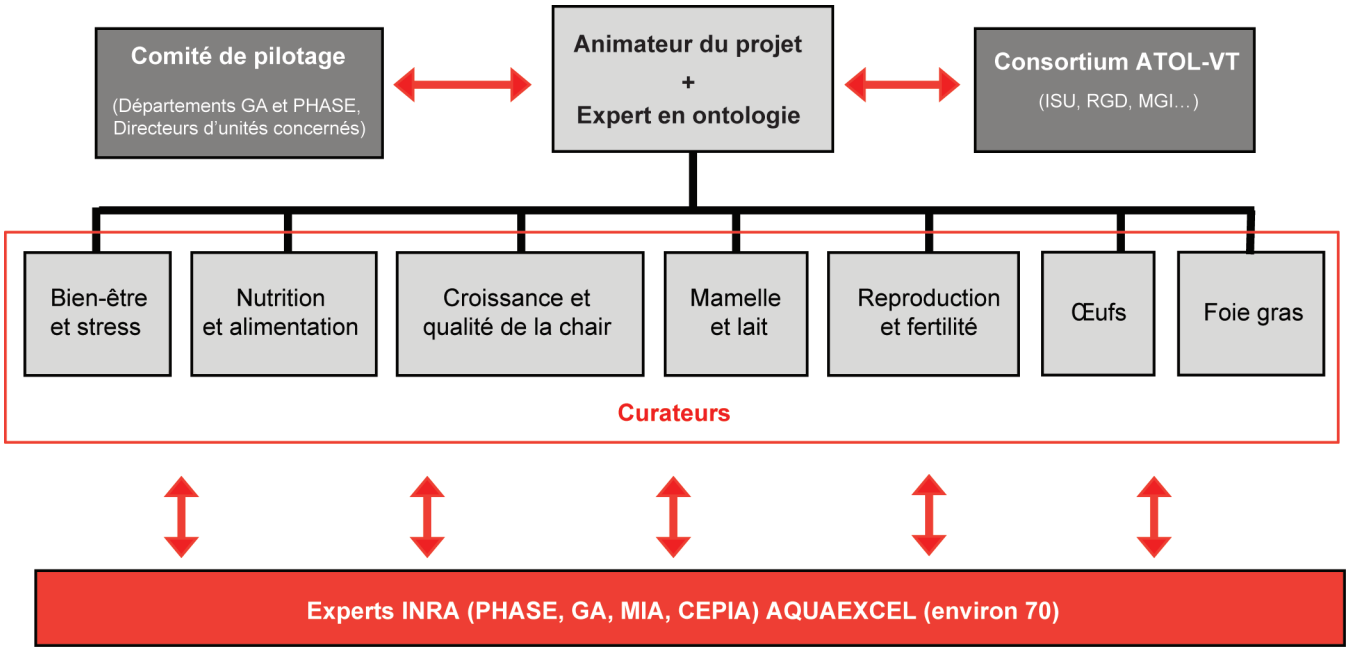
4 http://bioportal.bioontology.org/
5 http://www.informatics.jax.org/searches/MP_form.shtml
6 http://www.human-phenotype-ontology.org/
7 http://obofoundry.org/wiki/index.php/PATO:Main_Page
8 http://zfin.org/action/ontology/ontology-search
9 http://bioportal.bioontology.org/ontologies/VSAO?p=classes
10 Allèle : une des différentes versions d'un même gène à un même locus

dédié à la communauté travaillant sur les animaux d'élevage.

Une autre ambition était de construire une ontologie qui soit reconnue et partagée internationalement comme la référence dans le domaine des animaux d'élevage, avec comme corollaire, la possibilité de réutiliser l'ensemble des données publiées au niveau international dans le cadre de méta-analyses notamment. Plutôt que de développer un outil endogène risquant de n'intéresser que la communauté nationale, la partie française s'est associée au projet ATO pour « *Animal Trait Ontology* » en développement à l'ISU (« *Iowa State University* ») (Hughes *et al* 2008), sous les auspices de l'USDA- « *National Animal Genome* », qui visait des objectifs similaires. Ainsi, dès sa genèse, le projet s'est inscrit dans un cadre international en faisant reconnaître les spécificités des produits issus de l'élevage français. En plus de l'ISU et de l'Inra, en relation avec ses partenaires d'Agreenium, le consortium du projet ATO incluait le « *Rat Genome Database* » (RGD)[11] et le « *Mouse Genome Informatics* » (MGI)[12], deux consortiums américains ayant une finalité « santé humaine » et regroupant eux-mêmes diverses bases de données dédiées à ces deux espèces. Des partenaires européens, en particulier ceux du WUR (« *Wageningen University and Research* », Pays Bas)[13], au travers de contrats avec l'Union Européenne comme EADGENE (« *European Animal Disease Genomics Network of Excellence* »)[14] pour la santé animale ou - dans une moindre mesure - SABRE (« *Cutting Edge Genomics for Sustainable Animal Breeding* ») pour la reproduction bovine, ont également participé à ATO. Des représentants des filières animales comme l'EFFAB (« *European Forum of Farm Animal Breeders* »)[15] ont également montré leur intérêt pour un tel projet. Par la suite, le renforcement de l'ontologie pour la pisciculture s'est appuyé sur le réseau AQUAEXCEL[16], un projet « *Research Infrastructure* » de l'Union Européenne regroupant 17 partenaires de 10 pays différents, et visant notamment à la standardisation des caractères phénotypiques et de leurs mesures au sein des infrastructures aquacoles européennes.

Pour l'ensemble des partenaires, l'ontologie des caractères phénotypiques devait être la plus générique possible car devant couvrir des modèles de vertébrés très différents. Cette démarche était envisageable car nombre de fonctions et de régulations sont partagées au sein des vertébrés, même si des divergences très importantes sont observées au niveau de la morphologie (pattes/ailes/nageoires, par exemple), du mode de reproduction (ovipare/vivipare) ou du milieu de vie (aquatique/terrestre). Cette approche se situait donc en rupture avec les référentiels existants qui s'adressaient à une seule espèce, ou à un groupe d'espèces phylogénétiquement proches, entraînant la multiplication de caractères phénotypiques identiques et en limitant l'interopérabilité. Elle permettait aussi d'envisager la compréhension générique des mécanismes communs impliqués dans la construction de certains phénotypes. Si, pour l'Inra, l'ontologie devait concerner essentiellement les espèces d'élevage (bovins, ovins, porcins, volailles, poissons d'élevage…), elle n'excluait pas les espèces modèles (souris, rat, poissons zèbre, médaka…) faciles et peu coûteuses à élever, lorsqu'elles étaient pertinentes pour étudier les mécanismes d'élaboration des phénotypes intéressant l'élevage.

Figure 1. *Structure du réseau d'experts impliqués dans la construction d'ATOL.*



AQUAEXCEL : Réseau européen d'infrastructure aquacole d'excellence (FP7) ; CEPIA : Département de Caractérisation et Elaboration des Produits Issus de l'Agriculture de l'INRA ; GA : Département de Génétique Animale de l'INRA ; ISU : « *Iowa State University* » ; MGI : « *Mouse Genome Informatics* » consortium ; MIA : Département de Mathématiques et Informatique Appliquées de l'INRA ; PHASE : Département de Physiologie Animale et Système d'Elevage de l'INRA ; RGD : « *Rat Genome Database* » consortium ; VT : « *Vertebrate Trait ontology* ».

11 http://rgd.mcw.edu/
12 http://www.informatics.jax.org/
13 http://www.wageningenur.nl/en.htm
14 http://www.eadgene.info/
15 http://www.effab.org
16 http://www.aquaexcel.eu/

2 / La méthode suivie pour construire ATOL

2.1 / Mise en place d'un réseau d'experts

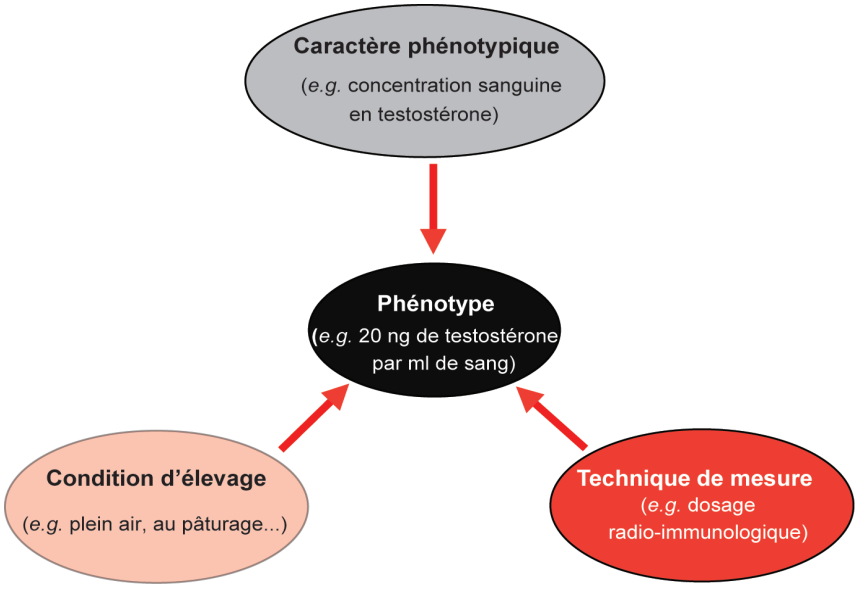
Les orientations et les finalités du projet ATOL ont été définies par un comité de pilotage regroupant des représentants des structures de recherche en physiologie et génétique animale de l'Inra. Le développement et la structuration de la base ont été confiés à un groupe de travail composé d'un animateur, d'un expert en ontologie dans le domaine de la santé humaine, et de cinq « curateurs ».

Les curateurs ont été chargés de la collecte, de la validation, de la définition des caractères phénotypiques et de leur organisation hiérarchique, avant leur introduction dans ATOL. Leurs compétences couvraient la croissance et la qualité de la chair, la nutrition et l'alimentation, la mamelle et le lait, la reproduction et la fertilité, le stress et et le bien-être animal et secondairement la production d'œufs et le foie gras (figure 1). Une cinquantaine d'experts, essentiellement issus de l'Inra, ont également été mobilisés par sous-groupes thématiques dans des réunions présentielles ou téléphoniques, comme force de proposition et de validation. Les experts ont été choisis en croisant leurs fonctions d'intérêt, leurs disciplines scientifiques et les différentes espèces d'élevage. Enfin, une équipe Inra spécialisée en terminologie, a rejoint le groupe pour compléter l'ontologie par une analyse sémantique de documents. L'ensemble de ce travail a été effectué en étroite connexion avec l'ISU afin de rendre l'ontologie ATOL compatible et complémentaire de l'ontologie VT (« *Vertebrate Trait ontology* »[17], Park *et al* 2013) que les chercheurs de cette université développaient au sein du projet ATO.

2.2 / Choix du périmètre d'ATOL

Dans la commande initiale, chaque caractère phénotypique devait être défini très précisément, en intégrant les facteurs environnementaux et les techniques de mesure, afin que la variation résiduelle du phénotype observé ne dépende plus que de la variabilité génétique. Cet objectif est apparu rapidement irréaliste. En effet, la combinaison de chaque caractère phénotypique avec la multitude de situations environnementales et de méthodes de mesure aurait conduit à la création d'un nombre considérable de concepts difficiles à générer et surtout très peu opérationnels. Nous avons donc posé

Figure 2. *Les domaines conceptuels requis pour construire un phénotype (génotype mis à part).*



l'hypothèse, partagée avec nos partenaires de l'ontologie VT (Shimoyama *et al* 2013), qu'un phénotype (hors fond génétique) pouvait être décrit à partir de 3 concepts (figure 2) : le caractère phénotypique (le caractère qui est mesuré), l'environnement dans lequel l'animal est élevé, et la méthodologie/technique employée pour mesurer le phénotype. Les deux premiers domaines conceptuels peuvent être définis dans le cadre d'ontologies, à savoir respectivement ATOL et EOL (« *Environment Ontology for Livestock* » pour Ontologie de l'Environnement des animaux d'élevage), cette dernière n'étant pas traitée dans le cadre de cet article.

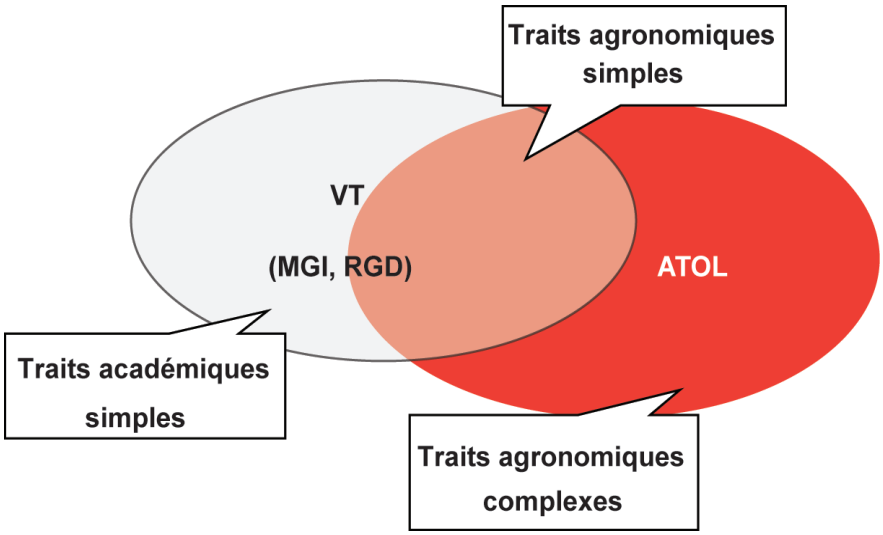
Concernant le troisième domaine conceptuel, les méthodologies de mesure d'un caractère phénotypique particulier peuvent être très nombreuses et en constante évolution. Il est donc utopique d'imaginer imposer une seule technique de référence à l'ensemble de la communauté œuvrant dans un même domaine Par exemple, pour mesurer la concentration sanguine en testostérone chez un veau, différents principes méthodologiques sont envisageables (chromatographie, spectroscopie, dosage immunologique…), mais au sein d'un même principe méthodologique (dosage immunologique par exemple) plusieurs techniques peuvent être déclinées (Dosage Radio-Immunologique (RIA) ou Immuno-Enzymatique (EIA), par exemple). Enfin, au sein d'une même technique (RIA, par exemple), les valeurs absolues mesurées dépendent de la procédure utilisée (volume de l'échantillon, extraction ou non de la molécule recherchée depuis le fluide vital, temps et température d'incubation…) et des réactifs (activité spécifique du traceur radioactif, nature et lot des anticorps, pureté des réactifs, qualité du standard de référence…). Ces différentes techniques peuvent donc conduire à des valeurs différentes du phénotype mesuré. Il nous a donc paru pertinent de nous référer à une base de protocoles méthodologiques référencés − qu'il reste à créer − à partir de laquelle des normalisations seront envisagées, dans le cadre d'analyses de données issues de méthodologies multiples.

2.3 / La démarche mise en œuvre pour construire ATOL

Dès le début de la réflexion avec nos partenaires américains, nous avons constaté que nos objectifs divergeaient partiellement. Si l'aspiration première à disposer d'une ontologie des caractères phénotypiques applicable aux vertébrés était partagée, nos partenaires américains voulaient l'organiser de manière académique (par fonction, organe, régulateur…) et ne prendre en compte que les caractères directement mesurables (taille, poids des gonades, pression artérielle…). Pour sa part, l'Inra voulait disposer d'une ontologie orientée vers les productions animales prenant en compte la qualité des produits mais aussi les caractères « calculés », ou « complexes », couramment utilisés dans ses programmes de recherche (taux de croissance, rapport gonado-somatique, % de refus alimentaires…). D'un commun accord, la démarche choisie a donc été de créer deux ontologies distinctes (VT et ATOL) tout en co-construisant les caractères d'intérêt commun (figure 3).

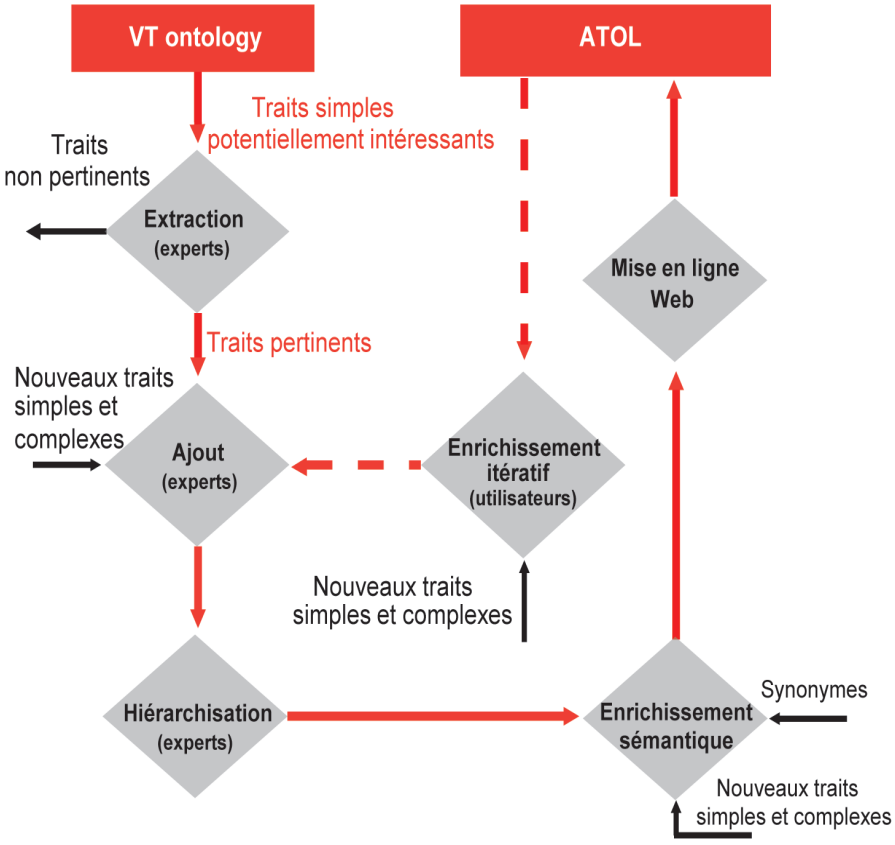
[17] http://purl.bioontology.org/ontology/VT

Figure 3. *Modalité du partage des caractères phénotypiques entre les ontologies ATOL et VT.*



Traits agronomiques simples : caractères phénotypiques d'intérêt pour les productions animales mesurés directement sur l'animal (exemple : poids vif de l'animal).
Traits académiques simples : caractères phénotypiques d'intérêt pour la recherche académique seule et mesurés directement sur l'animal (exemple : morphologie du muscle lisse).
Traits agronomiques complexes : caractères phénotypiques d'intérêt pour les productions animales exprimés comme une fonction mathématique plus ou moins complexe de caractères phénotypiques simples (exemple : le poids d'un organe peut être exprimé en % du poids du corps).
MGI : « *Mouse Genome Informatics* » ; RGD : « *Rat Genome Database* »

Figure 4. *Schéma récapitulatif des grandes étapes de la construction de l'ontologie ATOL.*



VT ontology : « *Vertebrate Trait ontology* » ; Trait : caractère phénotypique.

ATOL a été développée en six principales étapes. Dans un premier temps, nous avons procédé à une extraction automatique des branches[18] de VT (version du 06/02/2009) potentiellement pertinentes pour ATOL (caractères en lien avec les différentes productions animales). Dans un deuxième temps, et au sein de ces branches, les curateurs – avec l'aide de leurs experts – n'ont retenu que les caractères jugés pertinents, c'est-à-dire ceux expliquant pour partie la construction des phénotypes d'intérêt. Dans un troisième temps, les curateurs ont structuré ATOL en créant une hiérarchie orientée vers les productions animales (qualité/quantité) et le bien-être animal, et en cherchant à respecter le plus possible les niveaux d'intégration (molécule, cellule, tissu, organe, organisme…). Dans un quatrième temps, et pour chaque espèce animale, la hiérarchie et la liste des caractères avec leur définition ont été amendées et validées par le réseau d'experts. Dans un cinquième temps, les experts ont enrichi ATOL en proposant de nouveaux caractères phénotypiques. Dans un sixième temps, ATOL a été enrichie de nouveaux concepts et de nouveaux synonymes, à l'aide d'une analyse sémantique d'articles.

Dans la réalité, le processus a été moins séquentiel que présenté ci-dessus, il a fait appel à de nombreuses itérations entre travail des curateurs, des experts et apport de l'analyse sémantique (figure 4). Depuis la première extraction à partir de VT jusqu'à la mise en ligne de la première version d'ATOL, le processus a pris environ 3 ans, dépendant très fortement de la disponibilité des curateurs, des experts et du recrutement d'une personne sous contrat, entièrement dédiée au projet. Enfin, chacune des étapes a été menée en s'appuyant sur un certain nombre d'outils informatiques dont le rôle est présenté dans l'encadré 2.

3 / Les caractéristiques d'ATOL

3.1 / Evolution du nombre de caractères d'ATOL au cours de sa construction

a) Les caractères pertinents extraits de VT

La version du 06/02/2009 de VT comportait 4 182 termes, parmi lesquels 3 692 (88%) avaient une définition. Les curateurs ont sélectionné des branches potentiellement intéressantes pour un total de 1 625 caractères qu'ils ont soumis à leurs experts respectifs pour validation en tant que caractère « pertinent », « pertinent mais à modifier » ou encore « non pertinent ». A l'issue de cette phase, les caractères retenus dans ATOL (tableau 1) s'élevaient à 1 057, soit une

[18] Branches : premier niveau de l'arborescence de l'ontologie

Encadré 2. *Les outils informatiques mobilisés pour la construction de l'ontologie ATOL.*

Au cours des différentes étapes qui ont été suivies pour construire ATOL (« *Animal Trait Ontology for Livestock* »), plusieurs logiciels informatiques ont été utilisés pour faciliter le traitement en masse d'information ou pour rendre plus conviviale l'utilisation de l'ontologie.

L'étape d'extraction automatique des parties potentiellement pertinentes de l'ontologie VT (« *Vertebrate ontology* ») a été menée avec l'éditeur d'ontologie OBO-Edit(1) et avec des scripts écrits en langage de programmation Python. Chaque groupe d'experts, supervisé par un curateur, sélectionnait dans VT les caractères potentiellement intéressants. Les scripts Python récupéraient alors dans VT les propriétés de ces caractères (notamment les définitions textuelles) ainsi que leurs caractères parents(2) et enfants(3), organisant le tout en grandes branches thématiques, exportées en format Excel facilement utilisable par les experts.

La hiérarchie d'ATOL a été structurée pour les productions animales (dans un premier temps : Production laitière, Production de viande, Nutrition animale, Reproduction ; dans un second temps : Production de foie gras, Production d'œufs) et le bien-être animal en s'appuyant au maximum sur les niveaux d'intégration (molécule, cellule, tissu, organe, organisme…). Cette phase de sélection et d'organisation, qui reposait donc sur les compétences des experts pour le contenu, et celles des curateurs pour la structuration, a nécessité de nombreuses interactions - simultanées, en réseau - sur une même version de l'ontologie. Pour ces raisons, mais aussi pour pallier aux limites d'OBO-Edit, nous avons retenu ensuite l'éditeur d'ontologie *Protégé*(4) et son environnement collaboratif *Collaborative Protégé*(5).

La phase **d'enrichissement sémantique** a consisté à compléter ATOL avec des termes pertinents, extraits de corpus d'articles scientifiques par l'extracteur de termes BioYaTeA(6) (Golik *et al* 2013). Par exemple, à partir du texte « *methods to enhance milk fat yield and improve the fatty acid profile of milk fat* » (Elek *et al* 2008), l'outil BioYaTeA a extrait les cinq termes « *methods* », « *milk fat yield* », « *fatty acid profile* », « *milk fat* » et « *fatty acid profile of milk fat* ». Les trois termes soulignés sont pertinents pour ATOL, alors que « *methods* » et « *milk fat* » ne le sont pas. L'outil informatique FastR (Jacquemin 1999) a ensuite été appliqué au même corpus pour relier des synonymes à ces 3 termes et en découvrir de nouveaux. FastR utilise des règles de variation linguistique, morphosyntaxiques et sémantiques pour calculer de nouveaux termes à partir des termes d'ATOL. Le thésaurus généraliste WordNet (Miller *et al* 1990) fournit des classes de synonymes qui ont été utilisées par FastR pour générer des synonymes ATOL plus complexes. Par exemple, FastR propose le terme « *water intake* » comme synonyme du terme « *water consumption* » grâce à la synonymie entre « *intake* » et « *consumption* » définie par WordNet. Des ingénieurs de la connaissance, terminologues et experts, ont utilisé conjointement l'interface collaborative TyDI (« *Terminology Design Interface* ») pour valider et intégrer ces nouveaux termes dans l'ontologie (Nédellec *et al* 2010). Au-delà de l'extension des concepts de l'ontologie, cet enrichissement terminologique est particulièrement nécessaire dans l'objectif d'utiliser ATOL pour l'indexation automatique et sémantique des articles du domaine. L'ontologie doit donc intégrer la grande variété des termes utilisés (synonymes) pour identifier un concept donné dans un corpus.

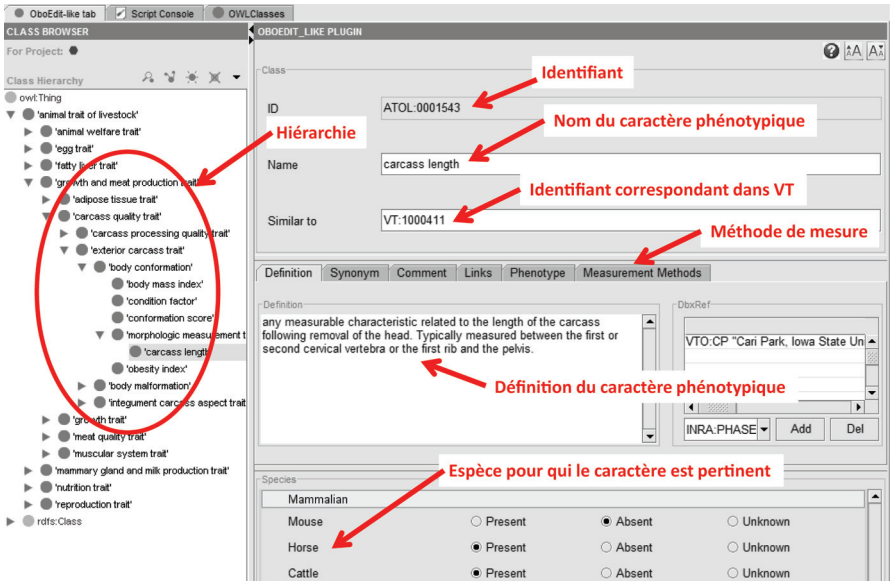
Enfin la **mise en ligne** d'ATOL, grâce entre autres à l'environnement WebProtégé, permet de visualiser, en plus de la hiérarchie, un ensemble d'informations destinées à définir le concept, à identifier son origine et à faciliter son utilisation (figure 5). Ce sont : *i)* un identifiant ATOL, complété par la référence de l'identifiant initial dans VT s'il y a lieu (pour « *carcass length* », ATOL:0001543, VT:10001411), en y associant la source du concept (groupe de travail INRA:PHASE ou issue d'une autre ontologie telle que « *Iowa State University Curator* »), *ii)* un nom (ici « *carcass length* ») qui correspond à l'usage le plus fréquent et les synonymes éventuels qui peuvent être exacts ou proches selon le degré de similitude fonctionnelle ou sémantique, *iii)* une définition dont la forme suit un cadre standardisé (par exemple : « *any measurable characteristic related to the length of the carcass following removal of the head. Typically measured between the first or second cervical vertebra or the first rib and the pelvis* »), *iv)* une validation par espèce de l'utilisation du caractère ATOL (par exemple : « absent » pour la souris, « présent » pour les bovins, porcins, ovins…), et éventuellement *v)* les liens vers des sites apportant des informations sur le caractère phénotypique (publications, gènes candidats, ARN, bases de données…), *vi)* les phénotypes connus associés au caractère ATOL, *vii)* les méthodes de mesure de ce caractère avec des liens vers des bases de données sur les procédures de mesure.

(1) http://oboedit.org
(2) Concept intégrateur immédiatement en amont (dans l'arborescence) du caractère auquel on se réfère.
(3) Concept immédiatement en aval (dans l'arborescence) du caractère intégrateur auquel on se réfère.
(4) http://protege.stanford.edu
(5) http://webprotege.stanford.edu/
(6) http://search.cpan.org/~bibliome/Lingua-BioYaTeA/

Tableau 1. *Nombre de caractères phénotypiques issus des branches potentiellement intéressantes (Initiaux) de l'ontologie VT* (« Vertebrate Trait ontology ») *parmi lesquels les caractères pertinents ont été retenus (Retenus de VT) pour la construction d'ATOL, ainsi que le nombre de caractères nouveaux ajoutés (Nouveaux) par les experts de l'INRA (version ATOL Juillet 2010).*

Caractères phénotypiques	Bien-être	Reproduction	Nutrition	Production de viande	Production laitière	Total
Initiaux	424	350	226	542	83	1625
Retenus de VT	250	234	106	419	48	1057
Retenus en % des initiaux	**60**	**75**	**47**	**77**	**58**	**67**
Nouveaux	37	93	135	106	381	752
Nouveaux en % retenus de VT	**15**	**35**	**127**	**25**	**794**	**71**

Figure 5. *Capture d'écran de l'éditeur d'ontologie Protégé présentant les principales informations contenues dans ATOL.*



La hiérarchie est organisée en branches (niveau N, par exemple, « *growth and meat production trait* »). Chaque branche est déclinée en sous-branches de différents niveaux (par exemple, N-1 ; « *carcass quality trait* » ; N-2 ; « *exterior carcass trait* » ; N-3…), jusqu'au niveau ultime (ici « *carcass length* »). Pour un caractère (ou trait) donné de niveau N, le caractère de niveau N+1 est appelé caractère phénotypique « parent », le ou les caractères de niveau N-1 étant appelé(s) caractère(s) phénotypique(s) « enfant(s) ».

diminution de 35%. L'élimination des caractères est liée à plusieurs motifs notamment leur nature trop éloignée des finalités des productions animales comme par exemple les caractères faisant référence à des pathologies (« nécrose du muscle squelettique » par exemple, qui peut être interprétée comme un phénotype (nécrosé) du caractère phénotypique « muscle squelettique »). D'autres caractères comme le « gras dorsal mesuré sur la dernière côte à 14 semaines » n'ont pas été jugés pertinents car d'une part, il ne s'agit pas d'une mesure usuelle en production porcine française et d'autre part, l'introduction de ce type de caractère composé (caractère mesuré × localisation × âge) aboutit à une explosion combinatoire des caractères présents dans l'ontologie et doit être envisagée plutôt en associant des caractères différents (gras dorsal, anatomie, âge à la mesure). Outre l'amélioration de la pertinence des caractères présents dans ATOL au regard des objectifs fixés, cette phase de sélection a permis de mettre en évidence l'importance d'un réseau d'experts par espèce pour appuyer ou non les choix des curateurs et aboutir à des définitions précises tout en restant génériques.

b) Les caractères ajoutés par les experts

Le tableau 1 présente le nombre de caractères nouveaux ajoutés dans ATOL par les experts Inra en plus des caractères issus de VT. Il apparaît fortement hétérogène selon les grandes branches. Ainsi les caractères relatifs à la « production laitière » ont fortement augmenté (+ 794%) car cette branche était peu renseignée par le consortium VT. D'autres branches - comme le « bien-être » et la « production de viande » - étant déjà bien étoffées, les nouveaux caractères proposés sont donc en proportion moindre (+ 15 et + 25% respectivement).

Après cette première étape appuyée sur l'ontologie VT, de nombreuses étapes de suppression et d'ajout de caractères ont été effectuées au cours de l'avancée du projet pour mieux prendre en compte leur généricité. Par exemple, pour la qualité des carcasses, il a fallu dissocier les caractères décrivant des mesures par espèce, afin de les regrouper en un caractère plus générique. Ainsi le caractère « *rib muscle weight* » valide pour le porc et « *thight weight* » pour les oiseaux ont été unifiés sous le caractère « *muscle weight trait* ». La combinaison de ce caractère avec une base anatomique propre à chaque espèce doit permettre de restituer ces caractères spécifiques. Des caractères calculés comme le « *carcass yield* » ont également été ajoutés. La figure 6 permet de visualiser le pourcentage de traits partagés, donc génériques, entre quatre espèces : le bovin, le porc, le poulet et la truite. La comparaison entre deux mammifères (bovin et porc) montre que les caractères relatifs à « *reproduction* », « *nutrition* », « *meat production* » et à « *welfare* » sont très fortement partagés (92 à 99%), alors que ceux ayant trait à « *milk production* » le sont plus faiblement (62%) car chez le porc, la qualité du lait à des fins de nutrition humaine est sans objet. Les caractères relatifs à « *meat production* » ou à « *nutrition* » (environ 90 et 80% respectivement), et dans une moindre mesure ceux relatifs à « *reproduction* » et à « *welfare* » (environ 40-70 et 60-70% respectivement) sont bien partagés entre mammifères, oiseaux et poissons. Seuls des caractères très spécifiques à certains groupes d'espèces comme « *milk production* » chez les mammifères, « *egg production* » chez les oiseaux et les poissons, ou « *fatty liver production* » pour les oies et les canards, ont un degré de généricité relativement faible.

Les experts européens du projet AQUAEXCEL ont secondairement validé

Figure 6. *Généricité des caractères phénotypiques (exprimé en %) au sein des 7 branches principales d'ATOL v6.0. Comparaison entre le bovin, le porc, le poulet et la truite.*

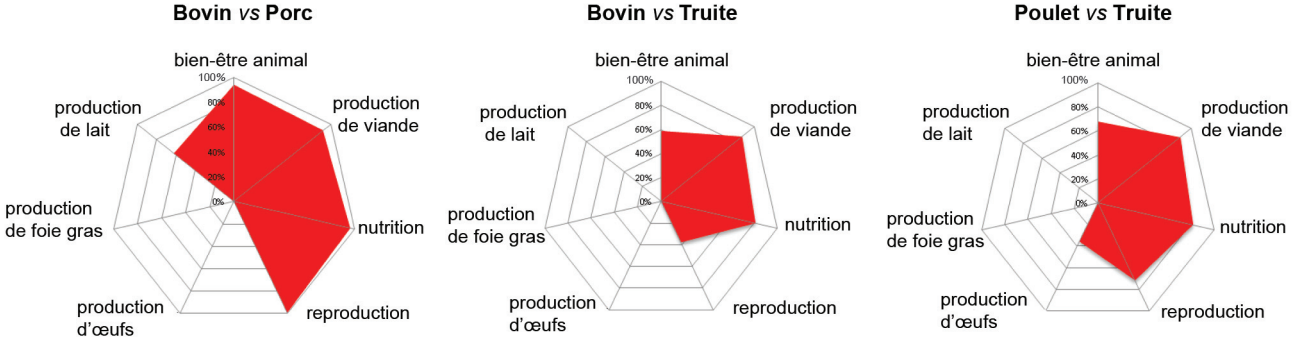
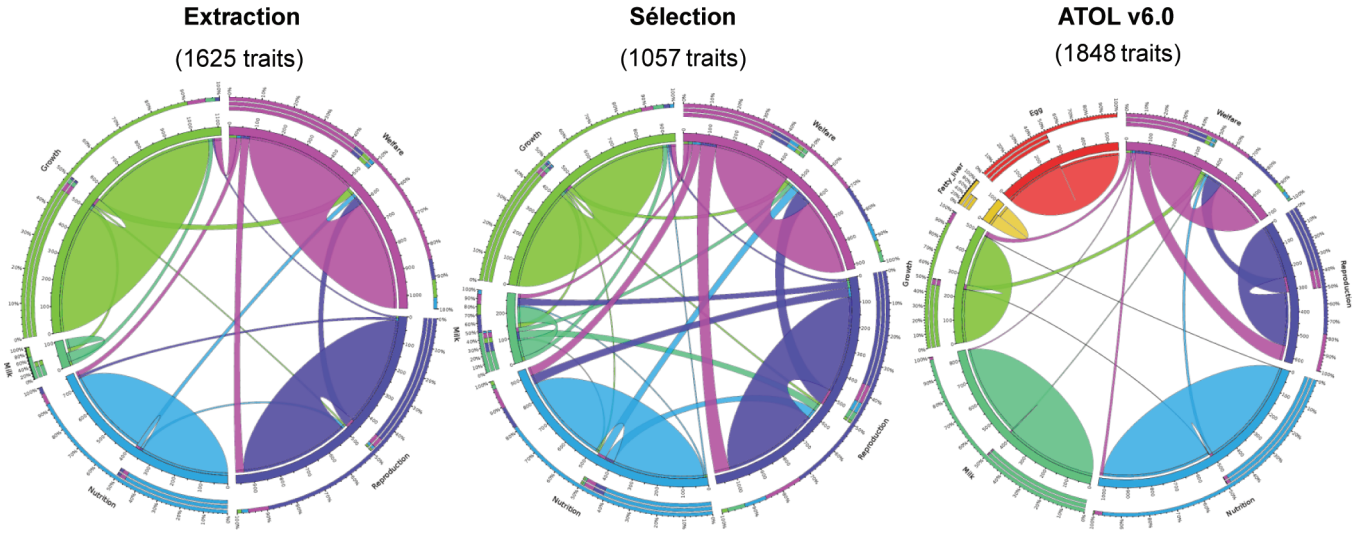


Figure 7. *Evolution du nombre total de caractères et du partage des caractères entre grandes branches au cours de trois étapes de la construction d'ATOL.*



Couleur des branches. vert pomme : « *growth and meat production* » ; vert menthe : « *milk production* » ; ocre jaune : « *fatty liver* » ; bleu : « *nutrition* » ; rouge : « *egg production* » ; magenta : « *animal welfare* » ; indigo : « *reproduction* ». Extraction (Caractères issus de branches de VT potentiellement intéressantes), Sélection (Caractères de VT conservés après expertise), ATOL v6.0 (version 6.0 d'ATOL après enrichissement par les experts et analyse sémantique).

les caractères existants chez les poissons et en ont ajouté de nouveaux. Ces différentes phases aboutissent aujourd'hui (ATOL version 6.0) à un total de 1 848 caractères (figure 7) assortis de leurs définitions et éventuellement de leurs synonymes. Dans cette version, le nombre des caractères phénotypiques est relativement équilibré au sein des grandes branches thématiques. Il varie autour de 19-24% pour les branches « *milk production* » et « *nutrition* », autour de 12-16% pour les branches « *meat production* », « *reproduction* », « *animal welfare* » et « *egg production* ». Seule la branche « *fatty liver* », très spécifique à la production française, reste en retrait (environ 2,5%).

c) Enrichissement d'ATOL grâce à l'analyse sémantique

Des analyses sémantiques effectuées sur le corpus de la revue « Animal » dans un premier temps, puis dans le cadre d'AQUAEXCEL sur le corpus du journal « Aquaculture », ont également permis d'enrichir ATOL de nouveaux concepts. Ainsi pour les poissons d'élevage, deux nouveaux caractères ont été introduits dans la branche « *animal welfare* » (« *amplitude ventilation* » et « *olfactory learning* »). Le nombre réduit de nouveaux concepts apportés par l'analyse sémantique montre que la couverture des caractères proposés par les experts était déjà très bonne. L'apport le plus important de l'analyse sémantique a été celui des synonymes. A partir de la revue « Animal », 156 synonymes ont été ajoutés à ATOL grâce à BioYaTeA et 171 grâce à FastR (Golik *et al* 2012). Ensuite, plus de 120 synonymes ont été ajoutés à partir de la revue « Aquaculture ». Ces ajouts ont notablement augmenté la couverture et la pertinence de l'ontologie pour l'indexation des articles.

d) Evolutions futures du nombre de caractères dans ATOL

Une ontologie n'est jamais figée et subit en permanence des mises à jour, ajouts, suppression de caractères en fonction par exemple de nouvelles branches à introduire dans l'ontologie, comme récemment celles sur la production de foie gras ou d'œufs. Les points de vue peuvent également évoluer avec l'arrivée de nouveaux experts de culture scientifique différente qui proposent de nouveaux caractères ou déclinent en caractères plus fins des caractères existants. C'est notamment le cas avec l'élargissement du consortium à de nouveaux partenaires européens. Par ailleurs, l'apparition de nouvelles technologies, soit plus précises, soit ouvrant de nouvelles possibilités d'exploration fonctionnelles (imagerie à haute résolution, puce SNP…), permettra sans doute de décliner certains caractères encore relativement « grossiers » en des concepts « enfants » plus pertinents et/ou plus proches du déterminisme génétique sous-jacent. Enfin, la mise en ligne en Avril 2012 de l'ontologie sur un site web dédié (http://www.atol-ontology.com/index.php/fr/), avec un formulaire de contact, permet un retour d'expertise de l'ensemble de la communauté scientifique. Dans les années à venir, il est probable que l'utilisation régulière de l'ontologie par les chercheurs, avec par exemple l'introduction des identifiants ATOL dans leurs publications, nécessitera l'ajout de nombreux traits, absents à l'heure actuelle. Ces ajouts devront donc être sans cesse contrôlés et validés par les curateurs afin de préserver l'homogénéité et la généricité souhaitées initialement.

3.2 / Organisation hiérarchique

Un effort particulier a consisté à organiser l'ontologie sous forme de hiérarchie de spécialisation, où chaque classe (ou une sous-branche de niveau N) est une spécialisation (ou une sous-classe, ou une classe enfant, ou sous-branche de niveau N-1) de ses superclasses. De plus, un seul type de relation entre concepts, « *is_a* », a été utilisé dans la version actuelle d'ATOL. Par exemple, « *adipose tissue fatty acid content* » (ATOL:0000074) et « *adipose tissue lipid oxydation* » (ATOL:0000075) sont deux sous-classes de « *adipose tissue lipid quality* » (ATOL: 0000073), les deux caractères enfants étant bien, l'un comme l'autre, des caractères qualitatifs des lipides du tissu adipeux. Ainsi, toutes les propriétés de la superclasse sont automatiquement vraies pour ses sous-classes (et les sous-classes de ces sous-classes). Grâce à cette notion d'héritage, la hiérarchie de spécialisation permet de simplifier la représentation des connaissances en factorisant tout ce qui est commun. Inversement, la hiérarchie de spécialisation permet d'automatiser le raisonnement en exploitant le fait que si une donnée est annotée par un caractère phénotypique de l'ontologie, on peut également considérer qu'elle est annotée par tous les ascendants de ce caractère (puisqu'ils sont plus généraux). Il devient alors possible de retrouver toutes les données non seulement en se référant

directement à « *adipose tissue fatty acid content* » ou « *adipose tissue lipid oxydation* » (respectivement ATOL:0000074 et ATOL:0000075), mais également en se référant indirectement à « *adipose tissue lipid quality* » (ATOL:0000073) qui est leur caractère parent. Ces deux aspects de simplification de la modélisation et d'automatisation du raisonnement sont possibles uniquement si les sous-classes sont des cas particuliers de superclasses. Il a donc fallu corriger les portions pertinentes de VT pour lesquelles ce principe n'était pas respecté.

Lorsque nécessaire, nous avons eu recours à l'héritage multiple en permettant à une classe (un caractère) d'avoir plusieurs parents directs. Ainsi, « *body weight* » (ATOL:0000351) est à la fois une sous-classe de « *animal performance trait* » (ATOL:0001516) et de « *growth trait* » (ATOL:0000855). En effet, les caractères à héritage multiple expliquent de manière toute aussi pertinente la construction de caractères phénotypiques appartenant à des branches différentes. Le partage des caractères entre les 7 grandes branches d'ATOL reste toutefois très inégal (figure 7). Les branches « *nutrition* », « *milk production* », « *fatty liver* », « *egg production* » partagent moins de 5% de leur caractères avec les autres branches. A l'inverse, la branche « *animal welfare* » partage plus de 30% de ses caractères avec les autres branches, en particulier celles des branches « *reproduction* » et « *meat production* », soulignant ainsi que le concept de « bien-être animal », au-delà de l'expression des comportements, permet une approche multicritère incluant une part importante de critères de production de nature zootechnique.

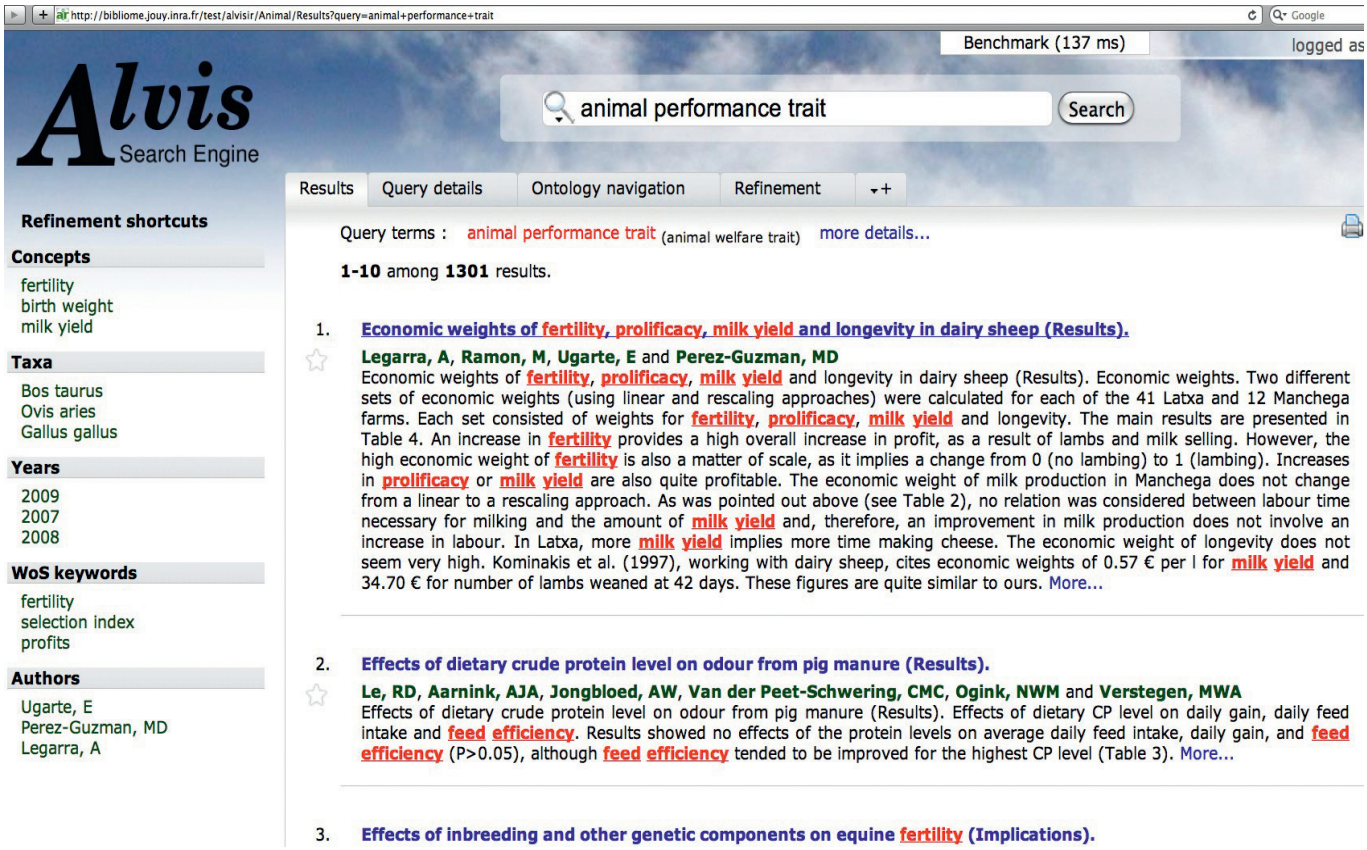
La dernière version d'ATOL, riche de ses nouveaux concepts et grâce à sa structure hiérarchique, permet une fouille bibliographique beaucoup plus performante qu'avec de simples requêtes par mot clé utilisées avec les moteurs de recherche de type Google. Par exemple, la requête « *animal performance* » avec le moteur de recherche sémantique *AlvisIR*[19] sur le journal « Animal » de Cambridge University Press entre 2007 et 2010, renvoie 1 301 résultats contre seulement 547 avec Google Scholar. En effet, le terme de la requête « *animal performance* » (figure 8) est interprété comme le concept de l'ontologie ATOL, c'est-à-dire aussi par tous ses caractères enfants et leurs synonymes (dont le nombre est supérieur à 40) définissant la performance, comme la fertilité, la production de lait, d'œuf, etc.

3.3 / Les difficultés rencontrées

a) Jusqu'où aller ou ne pas aller dans le choix des caractères

Si l'on souhaite une ontologie réellement opérationnelle et qui serve de référence à toute une communauté, il faut s'en tenir à son champ d'application et aux objectifs fixés. Une base trop fournie en caractères peu pertinents, mal orientée et mal structurée génère un outil trop complexe qui rebute les utilisateurs potentiels. La recherche de l'exhaustivité n'est donc pas souhaitable et constitue souvent un frein à l'élaboration d'un outil performant. De plus, le caractère

Figure 8. *Capture d'écran du moteur de recherche bibliographique AlvisIR montrant les résultats de la requête « animal performance trait » en utilisant ATOL.*
En utilisant le caractère phénotypique « *animal performance trait* », la requête effectuée sur l'ensemble de la collection des articles du journal «Animal» (CUP) prend en compte ce caractère et l'ensemble des caractères plus spécifiques parmi lesquels les termes en rouge (« *fertility, prolificity, feed efficiency...* »).



[19] http://bibliome.jouy.inra.fr/test/alvisir/Animal/

phénotypique d'intérêt ne correspond pas à une granulométrie particulière définie *a priori*. Par exemple, dans le cas de la qualité des tissus, le caractère peut se décliner tout aussi bien au niveau de la morphologie de la carcasse, de la tendreté de la viande, de la cellularité du muscle ou du pH musculaire. Dans ce cas, le niveau moléculaire peut aussi être pertinent puisque, par exemple, les concentrations d'IGF1 peuvent rendre compte de la prolifération des cellules musculaires, elle-même liée à la cellularité du muscle. Si les concentrations en ARN, et peut-être demain les marques épigénétiques des promoteurs de gènes myogéniques, peuvent parfois apporter des informations appropriées, il nous a semblé plus logique de nous en tenir à l'échelle de la protéine qui leur correspond et qui est généralement l'effecteur. Mais ce qui compte, c'est le lien explicatif avéré du caractère avec la qualité du tissu visé, la pertinence du caractère étant laissée à l'appréciation des experts du domaine. Nous n'avons donc pas suivi de règle générale, *a priori* plus satisfaisante conceptuellement, mais une démarche au cas par cas priorisant l'efficacité de l'ontologie.

b) Les caractères complexes ou calculés, pourquoi ?

Le consortium VT définit un caractère comme une caractéristique simple que l'on peut mesurer ou observer directement sur l'animal, un organe, un tissu ou même une cellule. De nombreuses variables d'intérêt agronomique sont en fait issues de calculs utilisant plusieurs caractères simples ramenés par exemple au poids vif de l'animal (rendement en carcasse, rapport gonado-somatique…). La question est donc de savoir si l'on peut considérer ces variables, que l'on appellera complexes ou calculées, comme des caractères à part entière. Dans un souci de comparaison entre espèces et de description fine des phénotypes au cours du développement et de la croissance de l'animal, il est essentiel de faire figurer ce type de caractères dans l'ontologie. L'information contenue dans un rapport gonado-somatique (poids des gonades ramené au poids du corps de l'animal et exprimé en %) est en effet plus pertinente que celle du simple poids des gonades qui, variant avec la taille de l'animal, ne donnera aucune indication sur son état de maturité sexuelle et ne permettra pas de comparaison entre individus. De même, le poids d'une carcasse de porc de 100 kg et d'une truite de 2 kg ne donne aucune idée du rendement en carcasse entre ces deux espèces qui peut être toutefois assez similaire (autour de 80%). *A contrario*, des animaux avec des poids de carcasse proches comme le lapin et la truite peuvent avoir des rendements très différents (60 *vs* 80%). La plupart de ces caractères calculés sont en fait des normalisations d'une mesure brute ramenée à un autre caractère ou paramètre pris en référence (poids ou taille de l'animal, surface, temps, nombre d'évènements…). A l'avenir, il nous faudra introduire dans la hiérarchie d'ATOL une relation supplémentaire de type « *normalization_of* » pour mieux rendre compte du positionnement de ces caractères. Ainsi, le caratère rapport gonado-somatique sera relié au caractère poids des gonades par la relation « *normalization_of* », ce qui permettra d'exploiter cette nouvelle propriété lors de fouilles bibliographies à l'aide d'ATOL.

c) Un nom de caractère concis ou précis ?

Le nom d'un caractère (ou étiquette du concept) est d'autant plus pertinent qu'il est précis, unique, univoque et couramment utilisé. C'est le cas par exemple du caractère phénotypique « *puberty* » (ATOL:0000431) qui correspond à l'âge auquel l'animal devient capable de se reproduire et de donner une descendance. En effectuant une fouille bibliographique pour ce caractère et pour une espèce donnée, la quasi-totalité des documents repérés traiteront effectivement de puberté. Mais nombre de caractères phénotypiques ne peuvent à la fois être concis et précis. Par exemple, si l'on veut sélectionner des documents traitant de « *milk fatty acid C10:0 concentration* » (ATOL: 0000644), une fouille avec ce nom étendu ne sélectionnera pas ou peu de documents traitant du sujet car la probabilité de trouver une chaîne de 5 mots ordonnés de la même manière est extrêmement faible. En revanche, le concept « *fatty acid C10:0 concentration* » permettra de repérer des documents beaucoup plus nombreux, avec un résultat de requête moins pertinent puisque les documents porteront sur la concentration en acide caprique dans différents tissus comme la viande, le foie, le sang, etc. Lors de la conception d'ATOL, nous avons recherché les noms de caractères les plus courts possibles. Mais dans certains cas, et pour gagner en précision, nous avons délibérément opté pour des noms composés de plusieurs mots (jusqu'à 5) en considérant que leur pertinence en termes de fouille bibliographique augmenterait au cours du temps au fur et à mesure que les auteurs des publications se réfèreraient aux termes univoques proposés par ATOL. Entre temps, il reste la possibilité de fractionner le nom du caractère – par les outils informatiques –, par exemple en « *concentration in milk* » et « *fatty acid C10:0* », pour obtenir un résultat de requête satisfaisant.

d) L'aspect temporel des caractères

La mesure de certains caractères est parfois associée au cycle de vie de l'animal comme son âge ou son stade de développement. Ainsi pour la qualité des carcasses, le caractère « épaisseur de gras dorsal à 14 semaines », tel qu'il est mesuré aux USA, ne correspond pas à une pratique en France et n'est pas standardisé au niveau international. L'introduction de caractères très associés au stade de développement expose au risque de devoir les décliner à différentes étapes de la vie de l'animal (épaisseur à 14 semaines, 5 mois, 6 mois…). Il est donc nécessaire de dissocier le caractère proprement dit des aspects temporels liés aux conditions de mesure. Comme pour les caractères liés à l'anatomie, c'est l'approche combinée (caractère + âge) qui a été la plupart du temps privilégiée (cf. 3.1/ b). Toutefois, cet aspect temporel a été conservé dans certains cas lorsqu'il peut constituer un caractère en lui-même. Par exemple, le concept « âge à l'abattage » peut être pertinent car si l'âge seul peut être traité comme une condition de mesure, l'âge à l'abattage est directement relié à la croissance de l'animal et peut donc être considéré comme un caractère, d'autant plus qu'il est ainsi générique pour toutes les espèces et partagé par l'ensemble de la communauté. Le même problème se pose pour les caractères liés au développement embryonnaire. Il a fallu en effet distinguer les caractères spécifiques du développement de ceux que l'on pouvait mesurer tout au long de la vie de l'animal. Par exemple dans la sous-branche « développement du muscle », le « nombre de fibres musculaires à la naissance » n'est pas spécifique puisque le caractère « nombre de fibres » peut être mesuré chez l'adulte comme chez l'embryon, en revanche « l'apparition de la première génération de myoblastes » est un caractère spécifique du développement embryonnaire.

e) Lien avec les paramètres environnementaux

Un autre problème posé dans la définition d'un caractère, et les limites qu'elle induit, est celui des conditions dans lesquelles ce caractère est mesuré. Si l'on s'intéresse aux capacités de croissance intrinsèque d'un animal, le gain de poids d'une truite élevée à 10°C n'apporte pas la même information que celui d'une truite élevée à 20°C. Outre l'effet classique positif de la température sur la vitesse de croissance des poïkilothermes, on peut se demander si l'on mesure bien les mêmes capacités de croissance entre ces deux conditions et donc le même caractère. Il apparaît toutefois irréaliste d'introduire dans la définition du caractère « taux de croissance » un lien avec la température d'élevage sans aboutir à une inflation du nombre de caractères liés à la multitude de combinaisons possibles, d'autant plus que la

variable est continue (poids à 10°C, poids à 20°C…). Il est donc nécessaire de faire abstraction de ce type d'information dans l'ontologie ATOL, tout en l'utilisant pour annoter les données de mesure à partir d'une autre ontologie (EOL) qui porterait cette fois sur l'ensemble des conditions d'élevage (facteurs biotiques et abiotiques).

Conclusion et perspectives

Le chantier ATOL (figure 4), qui a mobilisé fortement la communauté des scientifiques des départements PHASE, GA, CEPIA et MIA de l'Inra, mais aussi nos partenaires internationaux, débouche aujourd'hui sur un référentiel des caractères phénotypiques couvrant une très large partie des concepts utilisés dans le domaine des productions animales. Si l'ontologie est au cœur du projet, elle permet en outre – sous sa forme en ligne – de jouer un rôle de portail ouvrant des liens vers des sites porteurs d'informations complémentaires (structure des molécules impliquées dans les caractères, liste des phénotypes associés…). Dès à présent, ATOL permet de construire des bases de données – précises et interconnectables – sur les caractères phénotypiques. Le département PHASE l'utilise dans son projet TriPhase visant à répertorier très finement les recherches effectuées dans ses unités de recherche. Il est également envisageable de mettre à disposition des chercheurs un moteur de recherche bibliographique s'appuyant sur ATOL, aux performances bien supérieures aux classiques moteurs de recherche de type Google. L'ontologie devra toutefois évoluer en intégrant d'autres périmètres comme la santé animale, des productions plus modestes (laine, cuir…), ou s'élargir à d'autres espèces, pour autant que des spécialistes de ces domaines s'y impliquent. Pour que le dispositif de description des phénotypes soit totalement opérationnel, il doit être complété en décrivant les conditions d'élevage et de mesure. L'ontologie EOL (« *Environnement Ontology for Livestock* »), construite par l'Inra et ses partenaires européens du projet Aquaexcel, est déjà accessible. En revanche, la constitution d'une base de référence des protocoles de mesure est un chantier qu'il reste à mener et qui nécessitera à la fois une forte volonté et des moyens humains. Enfin, ATOL a été constituée pour et par des scientifiques. Cet outil prendra pleinement son sens et accroîtra son opérationnalité quand son appropriation par la communauté scientifique sera effective, par exemple en indiquant systématiquement dans les publications les identifiants et les noms standardisés des caractères phénotypiques proposés par l'ontologie, à la manière de ce qui est maintenant couramment pratiqué avec les noms scientifiques des espèces ou les séquences géniques de GenBank[20]. Mais au-delà de la communauté scientifique, les interprofessions disposent dès à présent d'un référentiel permettant d'organiser et de standardiser des bases de données phénotypiques dans le cadre, par exemple, de programmes de sélection génétique.

Remerciements

Les auteurs tiennent à remercier le département PHASE de l'Inra à l'initiative du programme ATOL et qui a soutenu financièrement dans le temps l'ensemble des réunions de travail et des équipements informatiques. Notre reconnaissance va également au comité de pilotage et aux différents experts qui se sont investis sans relâche dans ce projet et dont les noms sont consultables sur le site ATOL (http://www.atol-ontology.com/index.php/fr/les-acteurs-fr/experts/les-experts-fr). Merci également à James Reecy et Cary Park (USDA) pour les échanges très fructueux pour rendre compatibles ATOL et VT. Remerciements également aux acteurs du projet européen AQUAEXCEL (FP7 « *Aquaculture Infrastructures for Excellence in European Fish Research, Project number: 262336* ») qui a par ailleurs financé le salaire d'une personne employée en CDD durant deux ans. Enfin, nous remercions les lecteurs arbitres du journal qui nous ont permis d'améliorer la clarté du présent article.

Références

Bard J.B.L., Rhee S.Y., 2004. Ontologies in biology: design, applications and future challenges. Nat. Rev. Genet., 5, 213-222.

Blake J.A., Bult C.J., 2006. Beyond the data deluge: data integration and bio-ontologies. J. Biomed. Inform., 39, 314-320.

Bodenreider O., Stevens R., 2006. Bio-ontologies: current trends and future directions. Brief Bioinform., 7, 256-274.

Cimino J.J., Zhu X., 2006. The practical impact of ontologies on biomedical informatics. Yearb Med. Inform., 124-135.

Corcho O., 2006. Ontology based document annotation: trends and open research problems. Int. J. Metadata Semant. Ontologies, 1, 47-57.

De Rochambeau H., 2007. Les principes de l'amélioration génétique des animaux domestiques. C. R. Acad. Agric. Fr., 93, 1-9.

Elek P., Newbold J.R., Gall T., Wagner L., Husventh F., 2008. Effects of rumen protected choline supplementation on milk production and choline supply of periparturient dairy cows. Animal, 2, 1595-1601.

Golik W., Dameron O., Bugeon J., Fatet A., Hue I., Hurtaud C., Reichstadt M., Salaün M.C., Vernet J., Joret L., Papazian F., Nédellec C., Le Bail P.Y., 2012. ATOL: the multi-species livestock trait ontology. In: Proc. 6th Metadata and Semantics Research Conference, Springer Verlag Communications in Computer and Information Science Serie. Cadiz, Espagne, 289-300.

Golik W., Bossy R., Ratkovic Z., Nédellec C., 2013. Improving term extraction with linguistic analysis in the biomedical domain. Res. Comp. Sci., 129-143.

Groth P., Leser U., Weiss B., 2011. Phenotype mining for functional genomics and gene discovery. Methods Mol. Biol., 760, 159-173.

Hocquette J.F., Renand G., Levéziel H., Picard B., Cassar-Malek I., 2006. The potential benefits of genetics and genomics to improve beef quality. Anim. Sci. Papers and Reports, 24, 173-189.

Hocquette J.F., Boudra H., Cassar-Malek I., Leroux C., Picard B., Savary I., Bernard L., Cornu A., Durand D., Ferlay A., Gruffat D., Morgavi D., Terlouw C., 2009. Perspectives offertes par les approches en « omique » pour l'amélioration de la durabilité de l'élevage des herbivores. INRA Prod. Anim., 22, 385-396.

Hocquette J.F., Meurice P., Brun J.P., Jurie C., Denoyelle C., Bauchart D., Renand G., Nute G.R., Picard B., 2011. The challenge and limitations of combining data: a case study examining the relationship between intramuscular fat content and flavour intensity based on the BIF-BEEF database. Anim. Prod. Sci., 51, 975-981.

Hocquette J.F., Capel C., David V., Guéméné D., Bidanel J., Ponsart C., Gastinel P.L., Le Bail P.Y., Monget P., Mormède P., Barbezant M., Guillou F., Peyraud J.L., 2012. Objectives and applications of phenotyping network set-up for livestock. Anim. Sci. J., 83, 517-528.

Hu Z.L., Dracheva S., Jang W., Maglott D., Bastiaansen J., Rothschild M.F., Reecy J.M., 2005. A qtl resource and comparison tool for pigs: PigQTLDB. Mamm. Genome, 16, 792-800.

Hughes L.M., Bao J., Hu Z.L., Honavar V., Reecy J. M., 2008. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. J. Anim. Sci., 86, 1485-1491.

Jacquemin C., 1999. Syntagmatic and paradigmatic representations of term variation. In: Proc. ACL'99, 341-348.

[20] http://www.ncbi.nlm.nih.gov/genbank/

Miller G.A., Beckwith R., Fellbaum C.D., Gross D., Miller K., 1990. WordNet: An online lexical database. Int. J. Lexicograph., 4, 235-244.

Monget P., Le Bail P.Y., 2009. Le phénotypage des animaux : le nouveau défi ? Animal phenotyping: the new challenge. Renc. Rech. Rum., 16, 407-409.

Mungall C.J., Gkoutos G.V., Smith C.L., Haendel M.A., Lewis S.E., Ashburner M., 2010. Integrating phenotype ontologies across multiple species. Genome Biol., 11, R2.

Nédellec C., Golik W., Aubin S., Bossy R., 2010. Building Large Lexicalized Ontologies from Text: a Use Case in Indexing Biotechnology Patents. EKAW 2010. 514-523, Springer Verlag. Lisbon, Portugal, Oct. 11-15, 2010.

Park C.A., Bello S.M., Smith C.L., Hu Z.L., Munzenmaier D.H., Nigam R., Smith J.R., Shimoyama M., Eppig J.T., Reecy J.M., 2013. The Vertebrate Trait Ontology: a controlled vocabulary for the annotation of trait data across species. J. Biomed. Semantics, 4, 13.

Robinson P.N., Mundlos S., 2010. The Human Phenotype Ontology. Clin. Genet., 77, 525-534.

Shimoyama M., Nigam R., Sanders McIntosh L., Nagarajan R., Rice T., Rao D.C., Dwinell M.R., 2013. Three ontologies to define phenotype measurement data. Front. Genet., Vol 3, Art. 87, 10p.

Smith C.L., Goldsmith C.A.W., Eppig J.T., 2005. The Mammalian Phenotype Ontology as a tool for annotating, analysing and comparing phenotypic information. Genome Biol., 6, R7.

Teletchea F., Fostier, A., Le Bail P.Y., Jalabert B., Gardeur J.N., Fontaine P., 2007. STOREFISH: A new database dedicated to the reproduction of temperate freshwater teleost fishes. Cybium, 31, 227-235.

Watson R., Gee A., Polkinghorne R., Porter M., 2008. Consumer assessment of eating quality – development of protocols for Meat Standards Australia (MSA) testing. Aust. J. Exp. Agric., 48, 1360-1367.

Résumé

Les avancées technologiques récentes en biologie permettent la production de grandes quantités de données capables de décrire de plus en plus finement les phénotypes. Pour traiter en masse ces informations à l'aide de programmes informatiques et comparer les phénotypes provenant d'études différentes, il est indispensable de disposer d'un langage standardisé définissant sans ambiguïté les caractères phénotypiques auxquels pourront se référer des utilisateurs variés (généticiens, physiologistes, biochimistes, modélisateurs, producteurs…). L'absence d'un tel référentiel pour les animaux d'élevage a conduit l'Inra, en collaboration avec ses partenaires internationaux, à mettre en place une ontologie nommée ATOL (« *Animal Trait Ontology for Livestock* »). Celle-ci vise à définir les caractères phénotypiques des animaux d'élevage en les organisant en catégories, autour des performances (efficacité alimentaire, fertilité), des produits (production laitière, de viande, d'œufs, de foie gras) et des préoccupations sociétales (bien-être animal) se rapportant aux productions animales. Cet article explique les motivations à l'origine du projet, les objectifs poursuivis, la démarche adoptée et son originalité, ses limites et ses performances. Notre ambition est que cette ontologie, actuellement en accès libre sur la toile, soit largement utilisée pour référencer les caractères utilisés dans les publications et les bases de données, et ce pour favoriser une fouille bibliographique précise, facilitant ainsi l'intégration des informations à des fins de biologie systémique et prédictive.

Abstract

A controlled vocabulary for livestock phenotyping: the ATOL ontology

Recent technological advances allow the production of large biological datasets that makes the description of phenotypes more accurate. To analyze this huge amount of information with computers and thus compare phenotypes, it is essential to define a standard language that unambiguously defines phenotypic traits so as to serve as a reference, worldwide, to any possible user (geneticist, physiologist, biochemist, modeler, producer...). The absence of such a language/reference for livestock species has led Inra, in collaboration with its international partners, to develop an ontology that is called ATOL (*Animal Trait Ontology for Livestock*). Its aims are to define the phenotypic characters of livestock species, and allocate them to different types: performance traits (feed efficiency, fertility), production traits (dairy, meat, eggs, fatty liver) and societal traits (welfare). This article summarizes the objectives of the project, the original approach used to build the ontology, but also its current status and performance as well as its limitation. This ontology is publicly available on the web and expected to be widely shared worldwide for the common use of unique terms to annotate publications, databases or mine literature and thus promote systemic as well as predictive biology.

LE BAIL P.-Y., BUGEON J., DAMERON O., FATET A., GOLIK W., HOCQUETTE J.-F., HURTAUD C., HUE I., JONDREVILLE C., JORET L., MEUNIER-SALAÜN M.-C., VERNET J., NEDELLEC C., REICHSTADT M., CHEMINEAU P., 2014. Un langage de référence pour le phénotypage des animaux d'élevage : l'ontologie ATOL. In : Phénotypage des animaux d'elevage. Phocas F. (Ed). Dossier, INRA Prod. Anim., 27, 195-208.