# Un outil au service de la standardisation des bases de données : les ontologies ATOL/EOL

Marie-Christine Meunier-Salaün[1], Jérôme Bugeon[2], Alice Fatet[3], Isabelle Hue[4], Catherine Hurtaud[1], Claire Nédellec[5], Jean Vernet[6], Matthieu Reichstadt[6], Pierre-Yves Le Bail[2]

**Résumé.** Pour pouvoir traiter en masse les données phénotypiques issues de différentes expérimentations animales et différents sites d'élevages, il devient indispensable de rendre les bases de données interopérables en standardisant la dénomination des variables à l'aide de référentiels univoques et partagés par l'ensemble des utilisateurs (généticiens, physiologistes, biochimistes, modélisateurs, responsables de sites expérimentaux, animaliers, producteurs…). De tels référentiels ont été mis en place par l'INRA, en collaboration avec des partenaires nationaux et internationaux, sous la forme des ontologies ATOL (animal trait ontology for livestock) et EOL (environnement ontology for livestock). Leur contenu, leur structuration et leur utilisation, ainsi que les ambitions qu'elles suscitent sont rapidement évoqués dans le présent article.

**Mots clés :** bases de données, élevage, ontologie, standardisation, caractères phénotypiques

## Introduction

Un des objectifs actuels des chercheurs de l'INRA est de pouvoir prédire les performances des animaux d'élevage en trouvant les relations existantes entre les différents niveaux du fonctionnement du vivant (génome, épigénome, transcriptome, protéome, métabolome…) et les phénotypes d'intérêt pour la production (viande, lait, reproduction, foie gras…) ou le bien-être des animaux eux-mêmes (meilleure résistance aux maladies, meilleure capacité de réponses au stress, adaptabilité accrue) (Hocquette et al., 2012 ; Monget et Le Bail., 2009). Les chercheurs disposent pour cela de bases de données de qualité générées dans les sites d'expérimentation (IE (installations expérimentales) et UE (Unités Expérimentales)) de l'INRA. Elle sont la plupart du temps dédiées à une seule expérimentation (donc utilisant leur propre nomenclature) et manquent d'informations sur les conditions (métadonnées) dans lesquelles ces données ont été acquises (condition d'élevage, environnement, matériel de mesure…). Cette situation réduit considérablement la puissance potentielle des méta-analyses (démarche statistique combinant les résultats d'une série d'études indépendantes sur un problème donné). Il convient donc de standardiser les données pour les rendre comparables et les bases de données pour les rendre interopérables, et ce d'autant plus que le partage des données devient peu à peu obligatoire si elles ont été publiées ou financées par les pouvoirs publics nationaux et européens (OpenData). Cette standardisation est rendue possible en grande partie grâce aux ontologies comme ATOL[7] (ontologie des traits phénotypiques des

[1] *UMR PEGASE, INRA, Agrocampus Ouest, 35590 Saint-Gilles, France ; marie-christine.salaun@inra.fr, catherine.hurtaud@inra.fr*

[2] *LPGP, INRA, Campus de Beaulieu, 35000 Rennes, France ; pierre-yves.le-bail@inra.fr, jerome.bugeon@inra.fr*

[3] *UMR PRC, INRA, 37380 Nouzilly, France ; alice.fatet@inra.fr*

[4] *UMR BDR, INRA, ENVA, Université Paris Saclay, 78350, Jouy- en-Josas, France ; isabelle.hue@inra.fr*

[5] *INRA, UR1077 MIG, 78352 Jouy-en-Josas, France ; claire.nedellec@inra.fr*

[6] *Université Clermont Auvergne, INRA, VetAgro Sup, UMR Herbivores, 63122 Saint-Genès-Champanelle, France ; jean.vernet@inra.fr, matthieu.reichstadt@inra.fr*

[7] *7 http://www.atol-ontology.com/*

animaux d'élevage) et EOL[7] (ontologie des paramètres environnementaux d'élevage) (Golik et al., 2012 ; Le Bail et al., 2014).

Un phénotype (par exemple la valeur du poids de l'animal) résulte de la définition univoque de son **caractère phénotypique** (le poids entier de l'animal, le poids de la carcasse, de quel type de carcasse…), des **conditions dans lesquelles l'animal a été élevé** (température, luminosité, alimentation, cadre physique de l'élevage…) et de la **méthode/technique de mesure** (quelle unité de mesure, quel type de balance, quelle précision, à quelle heure de la journée, animaux à jeun ou non…). ATOL propose donc de préciser le caractère (ou trait) phénotypique auquel on se réfère, et EOL définit les différents paramètres environnementaux des élevages où sont menées les expérimentations. Pour les mesures, et les unités de mesure, il convient de se référer à des procédures standardisées qui, à terme, annoteront l'ensemble les caractères phénotypiques et paramètres environnementaux. Les deux ontologies s'adressent aux différentes espèces de rente (animaux de production), et sont organisées sous une forme hiérarchique pour s'y repérer plus facilement comme nous allons le voir ci-dessous.

## Comment est organisé ATOL

L'objectif du projet ATOL étant de disposer d'une ontologie orientée vers les productions animales, cinq grands thèmes ont été retenus. Ils associent une production à une fonction physiologique : l'adaptation et le bien-être, la nutrition et l'efficacité alimentaire, le système reproducteur et la fertilité, la glande mammaire et la production de lait et enfin la croissance et la qualité des carcasses et des viandes. A ces productions ont été ajoutées celles des œufs et du foie gras. La construction de cette ontologie a été initiée en 2009 par le Département PHASE (Physiologie animale et systèmes d'élevage), en collaboration avec des partenaires américains (Université IOWA, J. Reecy). Dans un premier temps, une analyse de l'ontologie VT (vertebrate trait) a permis d'identifier un premier ensemble de caractères associés aux productions animales et aux finalités considérées. Dans un second temps, de nombreuses étapes d'extraction de caractères, pertinents pour les objectifs définis, ont été menées à l'aide de logiciels d'extraction d'information sémantique dans des corpus bibliographiques. Cette démarche a permis d'ajouter de nouveaux caractères au premier ensemble, la totalité des caractères phénotypiques ayant ensuite été répartie de manière hiérarchique dans chaque finalité. Près d'une centaine d'experts a contribué à la constitution de groupes thématiques, pilotés par les curateurs, pour couvrir les différentes espèces de rente et les domaines scientifiques associés. Ils ont mené un travail de tri, de validation, de hiérarchisation et d'annotation des caractères phénotypiques. Ainsi chaque caractère phénotypique est constitué d'un **nom** « normalisé », en général celui qui est considéré comme le plus utilisé dans la littérature, et d'un **identifiant** unique sous la forme d'un URI (uniform resource identifier), par exemple ATOL:0001017 pour le caractère « meat colour ». Cet identifiant est essentiel. Son utilisation dans les bases de données de mesures phénotypiques, et dans le matériel et méthodes des articles scientifiques, permet d'avoir un repérage et une traçabilité précise des traits via la réalisation de requêtes informatiques univoques. Chaque caractère possède une **définition**. Des **synonymes** proches ou éloignés peuvent être associés au nom afin de tenir compte d'éventuelles variations syntaxiques (ex : meat colour et meat color) ou d'appellation comme pour le coefficient d'embonpoint (dans ATOL : Fulton index ou condition factor). Les synonymes permettent ainsi de regrouper sous un même caractère des concepts qui peuvent avoir des noms différents selon les laboratoires et pays mais qui correspondent pourtant au même caractère phénotypique. Chaque caractère est associé (via des annotations) à des espèces animales (espèces de rentes et espèces modèles) pour lesquelles le caractère en question est pertinent et mesurable. Des champs spécifiques, toujours sous forme d'annotations, permettent de faire référence aux auteurs du caractère et éventuellement à une autre ontologie dans laquelle ce caractère peut se retrouver de manière plus ou moins proche. Enfin, comme il est essentiel de considérer la méthode de mesure pour expliciter au mieux les valeurs phénotypiques, les caractères peuvent être reliés à des méthodes de mesures. Ces méthodes restent encore

peu renseignées dans l'ontologie, et se limitent le cas échéant à un renvoi à un protocole standard, disponible en format pdf via un lien ajouté dans les caractéristiques du caractère. Les spécificités de certaines méthodes de mesure étant très complexe (voir Le Bail et al., 2014), la constitution d'une ontologie des méthodes de mesure n'a pas été jugée pertinente.

L'ensemble de ces caractères a été intégré dans une **structure hiérarchique** allant du concept le plus général (caractères parents) au plus précis (caractères enfants). Les caractères parents et enfants sont reliés les uns aux autres par une relation de type « est un ». Par exemple, un caractère comme la couleur de la viande (caractère enfant) est un caractère faisant partie des caractères sur l'aspect de la viande (caractère parent), qui lui-même fait partie des caractères sur les qualités technologiques et organoleptiques de la viande (caractère parent plus général), qui font eux-mêmes partie des caractères sur la qualité de la viande (concept encore plus général), que l'on retrouve dans le thème dédié à la croissance et à la qualité des carcasses (**Figure 1**). Ainsi, toutes les propriétés de la classe supérieure (par exemple « qualité de la viande ») sont automatiquement vraies pour ses sous-classes et leurs niveaux inférieurs (par exemple « qualités technologique et organoleptique de la viande » etc.).


***Figure 1**. Exemple d'organisation hiérarchique du trait « meat colour » (couleur de la viande).*

Un caractère peut toutefois avoir deux parents lorsque ce caractère est pertinent pour deux thèmes différents. On parle alors d'héritage multiple. 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 septembre 2017, l'ontologie ATOL contenait au total 2450 caractères, répartis entre les différentes branches comme indiqué dans la **Figure 2** dans la page suivante.

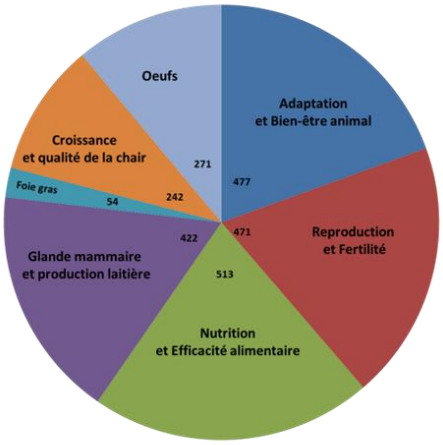


***Figure 2.*** *Répartition du nombre de traits d'ATOL par grands thèmes (branches) en 2017.*

## Comment est organisée EOL

Outre la description du trait phénotypique par ATOL, la compréhension d'un phénotype passe également par une caractérisation précise de l'environnement d'élevage qui influe directement sur l'animal et ses performances. L'ontologie EOL a été créée *de novo* pour décrire, de manière générique, les systèmes d'élevage et les conditions environnementales. On ne parlera plus ici de caractères mais de paramètres qui vont expliciter l'environnement physique, chimique et biologique de l'animal. La construction de l'ontologie EOL a été initiée par un groupe pilote d'une dizaine de personnes au sein du projet européen d'infrastructures « AQUAEXCEL » portant sur les milieux aquatiques, en s'appuyant sur la littérature adéquate y compris certains référentiels. L'ontologie à ensuite été élargie au milieu terrestre pour inclure l'ensemble des espèces de rente, en écho à la construction de l'ontologie ATOL.

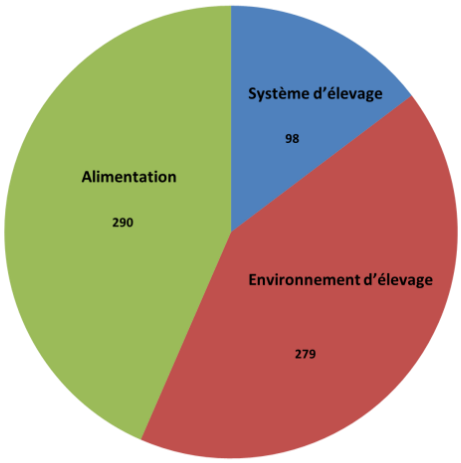


***Figure 3.*** *Répartition du nombre de paramètres d'EOL par grands thèmes en 2017.*

En septembre 2017, l'ontologie EOL contenait 637 paramètres (**Figure 3**) répartis de manière hiérarchique en trois branches principales : 1) le système d'élevage qui explicite brièvement comment et pour quelles raisons le système d'élevage a été conçu, 2) l'environnement d'élevage qui décrit les conditions biologiques, chimiques et

physiques dans lesquelles se trouve l'animal, 3) l'alimentation, qui décrit la qualité et les conditions de distribution de l'aliment.

Chaque paramètre de l'ontologie EOL est constitué des mêmes propriétés que celles des caractères ATOL, comme le nom, l'URI (par exemple EOL:0000001), la définition, les synonymes… Seules les notions d'espèce ou de race n'apparaissent pas ici. Elles sont néanmoins « annotables » par des ontologies dédiées : *NCBI Taxonomy* pour les organismes, *Livestock Breed Ontology* pour les races.

La combinaison des caractères phénotypiques (ATOL) et des paramètres des conditions d'élevage (EOL) permettra une annotation puissante des bases de données phénotypiques animales avec des métadonnées explicites, prérequis indispensables à des méta-analyses et des recherche sémantiques plus performantes. C'est déjà le cas au sein du consortium international FAAG pour annoter les génomes animaux (Tuggle et al., 2016).

## Comment utiliser ATOL/EOL

Les deux ontologies sont disponibles à l'adresse suivante : http://www.atol-ontology.com. Ce site internet présente le projet, les acteurs et partenaires ayant contribué à la création d'ATOL et EOL, et permet d'accéder aux ontologies ATOL et EOL et éventuellement de les télécharger. L'utilisateur peut visualiser indépendamment les deux ontologies du projet en suivant les liens dans l'en-tête des pages (Les Ontologies>Visualisation) et en cliquant sur le nom de l'ontologie dans la barre verte.

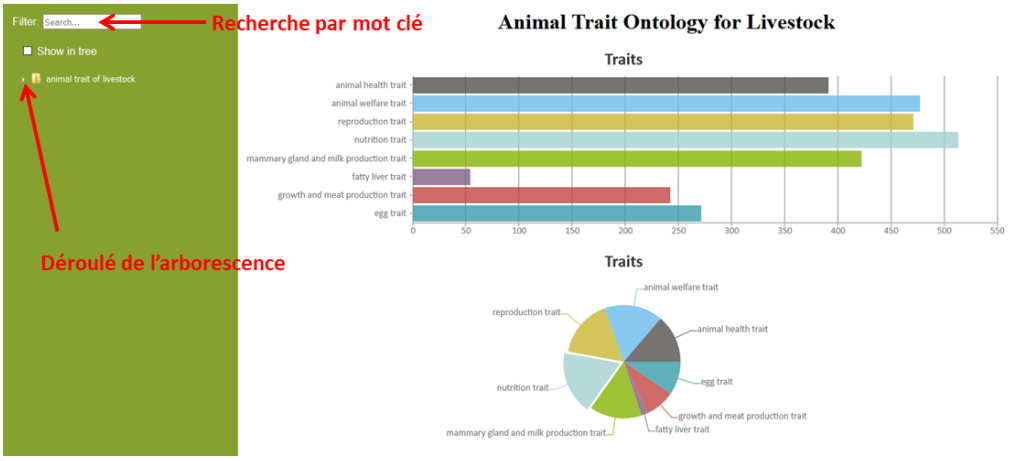


***Figure 4**. Page d'accueil de l'ontologie ATOL.*

Sur cette page, qui présente aussi les statistiques de l'ontologie (**Figure 4**), il est possible d'effectuer des recherches dans l'arborescence des ontologies afin de visualiser un caractère/paramètre particulier selon deux manières :

- ✓ soit en tapant un mot clé dans la fenêtre du filtre,
- ✓ soit en cliquant sur le petit triangle en face de chaque item afin de dérouler étape par étape l'arborescence selon le sujet d'intérêt de l'utilisateur. Lorsqu'on clique sur un item, les informations associées (nom, identifiant, synonymes, définition…) deviennent ainsi visibles.

Enfin, il est possible de télécharger les différentes ontologies au format OWL, afin de les utiliser dans des logiciels spécialisés tels que Protégé (https://protege.stanford.edu/).

## Un exemple d'annotation d'une base de données recomposée

L'exploitation d'un corpus de données phénotypiques pour réaliser des méta-analyses ne peut être automatisée sans une annotation homogène des caractères dans chacune de ces bases. Par exemple, le schéma de la **Figure 5** décrit la construction d'une base de données unifiée à partir de deux expériences menées séparément dans deux laboratoires et portant sur l'effet de la température de l'eau (paramètre de l'environnement) sur le facteur de condition (trait phénotypique) des truites âgées d'un an, ces deux éléments étant référencés dans chacune des ontologies en prenant en compte les synonymies (facteur de condition = condition factor = K = Fulton index = coefficient d'embonpoint).

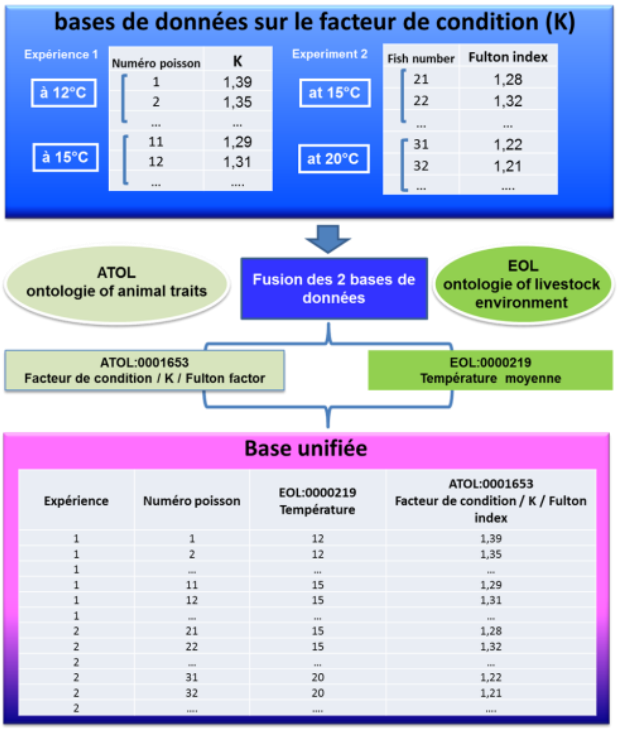


***Figure 5**. Exemple de fusion de deux bases de données issues de mesures du coefficient de condition (ou coefficient d'embonpoint), en fonction de la température de l'eau du bassin, chez la truite arc-en-ciel.*

Cette démarche permet de créer automatiquement un fichier de données unique et homogène en sélectionnant dans chaque expérience 1) les données associées aux valeurs des facteurs de variation, ici la température (EOL:0000219) qui varie de 12 à 20 °C, et 2) celles associées à la variable mesurée « facteur de condition » (ATOL:0001653), reconnues comme équivalentes (K, Fulton index) au sein des deux expérimentations.

## Discussion-conclusion

Construire une base de données en s'appuyant sur les référentiels ATOL et EOL présente donc de nombreux avantages en termes de partage de données au niveau institutionnel, national ou international (les deux ontologies ont été co-construites avec des partenaires européens et américains), ce qui explique leur présentation en anglais. Nous sommes bien conscients qu'une version française doit être mise en place si l'on veut accroître son utilisation par nos dispositifs expérimentaux, en particulier avec le développement de l'élevage de précision. Cette version française devrait être effective fin 2018. L'intégration des nombreuses métadonnées accompagnant les phénotypes expérimentaux peut sembler fastidieuse *a priori*, mais elle confère à la base de données une valeur décuplée en permettant sa réutilisation de multiples fois, longtemps après sa conception, au fur et à mesure de l'apparition de nouveaux questionnements scientifiques. Pour cette raison, le projet pilote D-ONT (2016-2018) de l'INRA a engagé une réflexion pour relier entre elles ses différentes bases de données sur

l'animal en utilisant ATOL et EOL. L'élaboration d'une application permettant d'accéder directement aux items des ontologies lors de la construction du canevas de la base de données serait idéale et devrait lever un certain nombre de réticences. Cette démarche est en cours d'élaboration dans le cadre de Cati (Centre automatisé de traitement de l'information) de l'INRA et en particulier le Cati Sicpa santé (Systèmes d'informations et calcul pour le phénotypage animal ciblé sur la santé), cette structure ayant pour mission d'unifier et homogénéiser l'acquisition et l'ensemble des bases de données expérimentales sur les animaux d'élevage de l'INRA. Un effort important devra également être consenti pour renseigner la base des procédures méthodologiques afin de donner toute sa cohérence au dispositif de saisi des données phénotypiques.

Toutefois, les ontologies sont des outils vivants qui évoluent en fonction des besoins de chaque utilisateur et de l'apparition de nouveaux caractères phénotypiques générés par l'accessibilité à de nouvelles technologies (imagerie fonctionnelle en temps réel…). Dans le cadre du projet AHOL (animal health ontology for livestock, financement par le métaprogramme GISA), un travail a été engagé par l'INRA pour étendre le périmètre d'ATOL à la santé animale. Tout utilisateur peut à tout moment contacter un des gestionnaires d'ATOL via le site web[7] pour proposer un nouvel item qu'il ne trouverait pas et dont il aurait besoin pour renseigner sa base de données. ATOL et EOL doivent être au service des agents INRA du domaine animal, et construites pour eux et par eux, cette condition étant indispensable pour permettre l'utilisation en routine, l'évolution et la pérennité de ces référentiels.

## Références bibliographiques

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: proceedings of The 6th Metadata and Semantics Research Conference (MTSR 2012), pp 289-300. *Springer Verlag Communications in Computer and Information Science Serie*. Cadiz, Espagne, 28 au 30 novembre.

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

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, Nédellec 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'élevage. Phocas F. (Ed). *Dossier, INRA Prod. Anim.*, **27 :**195-208.

Monget P, Le Bail P-Y (2009) Le phénotypage des animaux : le nouveau défi ? Animal phenotyping : the new challenge? In : Proceedings of the « 16è Rencontres autour des Recherches sur les Ruminants ». Institut de l'Élevage - INRA, Paris, pp 407-409.

Tuggle CK, Giuffra E, White SN, Clarke L, Zhou H, Ross PJ, Acloque H, Reecy JM, Archibald A, Bellone RR, Boichard M, et al. (2016) GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. *Anim Genet.* **47 :** *528-533.*