
BIF-BEEF

Construction et règles d'usage d'un gisement de données au service de la biologie intégrative.

«ceux qui seront capables d'intégrer [...] images, données génomiques, phénotypiques, économiques et environnementales, seront les acteurs et les détenteurs de la connaissance du vivant» (Florian GUILLOU, 23 mars 2009)

La Recherche sur la production de viande bovine est une activité coûteuse en temps et argent, en raison de la durée des cycles biologiques et de la valeur marchande de ces animaux. Par ailleurs, et au contraire des travaux en génomique, il n'existe pas de techniques « haut débit » pour produire de la connaissance sur le phénotype bovin.

Cependant, à partir des travaux terminés ou en cours, il est possible de réconcilier des informations à « haut volume », pour lesquelles la structuration apporte un intérêt nouveau et obligatoire dans la perspective d'améliorer la connaissance de lois biologiques générales et d'envisager la modélisation prédictive.

L'intérêt du regroupement méthodique et généralisé de toutes les données du phénotype bovin est perçu par chacun des producteurs d'information. Cependant il existe des obstacles objectifs à la libération des chiffres : propriété morale du producteur, validation des chiffres, compréhension, statut juridique de la donnée, équilibre des contributions et des partenariats,

De nombreuses initiatives voient le jour en 2009 pour faciliter le « data sharing ». La présente note n'ambitionne pas d'apporter une réponse définitive à chacune des questions que ces projets suscitent, néanmoins, elle mentionne quelques règles qui jusqu'à nouvel ordre, seront respectées par les acteurs URH.

Historique

Le programme transversal « muscle-viande », animé par C2M (B. Picard) et terminé en 2006, a été le moment fondateur pour l'élaboration d'une première base de données interne à l'Unité (FiLiCol). Celle-ci rassemble des données concernant l'élevage et la croissance des animaux (en lien avec ALADIN), les caractéristiques d'abattage, la composition corporelle des carcasses, les propriétés des muscles et de la viande d'expérimentations coordonnées par l'URH. D'autres données structurées pour les besoins d'un travail post-doctoral (Nicola SCHREURS) ont été capitalisées dans cette base par l'apport de 3 mois de MOO, payés par C2M.

A l'initiative de J.F. Hocquette, d'autres sources d'information sont venues enrichir la base FiLiCol (Gemqual, Qualvigene, Mugene), cette fois par l'apport d'un CDD de 16 mois financé par un contrat région (Lifegrid). A ce niveau la base est devenue un outil significatif pour les études en « Biologie Intégrative et Fonctionnelle » (BIF-beef).

Avec l'arrivée d'un ingénieur informaticien à temps plein, le travail méthodique de collecte, structuration et validation va entrer en régime continu. Il s'agit autant de valoriser le passé, que de gérer le présent, et de préparer l'avenir. Dans le futur proche sont attendus à C2M (début 2010), un thésard et un CDD post doctoral dont le travail va reposer sur les données de la base BifBeef.

Cette base compte aujourd'hui 20 espèces bovines, males, femelles et castrés, totalisant 5000 animaux et 600 000 mesures.

Un échange récent avec D. Durand nous permet également d'identifier de nombreuses données candidates : Contrats CIV, LIPIVIMUS, PROSAFEBEEF, OFIVAL-CUISSON, DEFIVIANDE, xxx-AQS, xxx-ADIV....

Du fait des projets de recherche, de l'ambition PHASE sur le phénotypage à haut débit, de l'actualité de ces données, du travail d'élaboration d'un dictionnaire des mesures (ontologie ATO), de la prise de fonction de Matthieu REICHSTADT, il est essentiel d'engager résolument le travail d'agrégation. Il faut noter par ailleurs que ce travail est d'autant plus facile que chacun garde à l'esprit la nature des mesures et les conditions de leur production.

Les producteurs de données

Pour de nombreuses raisons, en particulier de traçabilité, il est essentiel de pouvoir associer à chaque chiffre, un opérateur et un responsable. L'opérateur est le plus souvent le technicien qui a produit le chiffre, le responsable est l'auteur du protocole.

De même les grands groupes de données doivent être facilement identifiables : FILICOL, NICOLA, GEMQUAL

Ces 3 marqueurs : groupe de données d'expérimentation, opérateur, responsable sont stockés dans la base et associés à chacune des mesures.

Enfin, tant pour des raisons juridiques que morales, les données ne doivent pas être rendues publiques de façon incontrôlée. Chaque donnée est donc par défaut placée sous embargo. La date de levée d'embargo est fixée par le responsable de la mesure (il est souhaitable de ne pas poser d'embargo anormalement long, 6 ans semble un maximum, après la fin du protocole).

Les informaticiens

Ils sont chargés de la mise en forme des données, de l'identification précise des variables selon le référentiel des mesures, de l'incrémentation de la base et de la sécurité des données.

Par ailleurs, ils doivent prendre en compte la validation méthodique des chiffres et s'assurer par des traitements systématique qu'une requête ou un copier-collé ne sont pas à l'origine d'erreur massives et de données erronées.

A ce titre les informaticiens doivent suspecter les chiffres, se les approprier et leur donner du sens pour en garantir la valeur et accréditer leur auteur en le protégeant des erreurs de transport toujours à redouter.

Ils s'interdisent l'extraction et la fourniture de données qui proviendraient de demandes non coordonnées par l'instance évoquée ci-après.

Les utilisateurs

Le contenu de la base BIF-BEEF est placé sous le contrôle d'un Comité Scientifique d'Utilisateurs (CSU), de 4 membres permanents (B. Picard, J.F. Hocquette, D. Durand, D. Micol). A leur initiative, ce CSU sera ponctuellement élargi à d'autres responsables de protocole.

Le rôle de ce CSU est :

- De promouvoir le gisement de données, en suscitant la livraison de données candidates,
- De prendre connaissance régulièrement de la typologie du contenu,
- De statuer sur les demandes et projets qui s'appuieraient sur ces données,
- De s'assurer que le responsable des mesures soit informé de l'utilisation de ses chiffres,
- De recueillir l'autorisation, l'avis et les conseils si besoin de ce responsable
- De choisir, suivant les cas, la forme de la reconnaissance intellectuelle envers le responsable (co signatures, remerciements...)
- De veiller au respect des délais d'embargo.

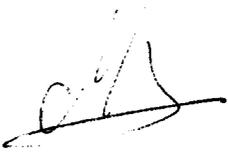
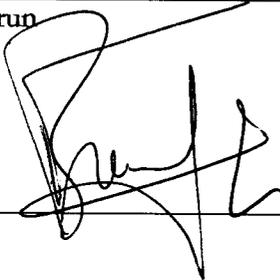
Et d'une façon générale, d'élaborer toute disposition de nature à permettre au mieux la valorisation du contenu, sans qu'il ait de préjudice réel ou supposé pour les fournisseurs de données, aux sens déontologique.

C'est sur l'ensemble de ces bonnes pratiques que doit se construire la confiance entre les acteurs concernés, par la même la pérennisation indispensable du travail d'agrégation, et sur lesquelles s'engagent les signataires ci-après.



Ce texte restera en vigueur à l'URH aussi longtemps qu'un autre texte de portée plus générale ne sera pas venu s'y substituer. En effet, le Département PHASE confie actuellement à un groupe de personnes la rédaction d'un projet ayant le même objet (voir la commande en annexe).

URH, Theix, le 25 Février 2010

B. Picard 	D. Durand 
D. Micol 	J.F. Hocquette 
M. Reichstadt 	J.P. Brun 

ANNEXE

----- Original Message -----

From: "Philippe Chemineau" <philippe.chemineau@tours.inra.fr>
To: "Poupon Anne" <Anne.poupon@tours.inra.fr>; "Brun Jean-Paul" <Jean-Paul.Brun@clermont.inra.fr>; "Vignon Xavier" <Xavier.Vignon@jouy.inra.fr>; "Hennequet-Antier Christelle" <christelle.hennequet@tours.inra.fr>; "VanMilgen Jaap" <Jaap.Vanmilgen@rennes.inra.fr>
Cc: "Guillou Florian" <Florian.Guillou@tours.inra.fr>; "Malpaux Benoît" <Benoit.Malpaux@tours.inra.fr>; "Coulon Jean-Baptiste" <jean-baptiste.coulon@clermont.inra.fr>; "Leroux Dominique" <dominique.leroux@tours.inra.fr>
Sent: Sunday, December 20, 2009 3:32 PM
Subject: Partage des donnees et bases de donnees

Bonjour,

Je reviens avec retard sur les conclusions de notre discussion du 27 octobre dernier en réunion de DUs de PHASE sur le partage des données et les conditions d'accès et d'utilisation des bases de données, pour vous proposer (AP, JPB, CHA et JVM) de réfléchir à deux points concernant ce sujet:

(1) Rédiger une "charte" de partage des données dans PHASE, qui abordera les questions liées à la "propriété" des données, aux conditions d'accès (délai, accessibilité, etc), à leur utilisation et leur valorisation notamment dans le cas de méta-analyses, à leur protection (dépôt, etc), et aux collaborations éventuelles hors PHASE (GA) ou hors INRA (INIST CNRS).

(2) Imaginer des propositions de mesures incitatives que la direction de PHASE pourrait prendre afin de stimuler le dépôt des données dans des bases.

Toute autre suggestion sera la bienvenue.

Je propose que vous aboutissiez à la fin du premier trimestre 2010, pour une nouvelle présentation en réunion des DUs de PHASE (mais c'est la nouvelle direction de PHASE qui en décidera).

Merci pour votre aide dans ce dossier très important pour le développement de la biologie intégrative.

Bien cordialement.

Philippe Chemineau
INRA
Chef du Département
"Physiologie Animale et Systèmes d'Élevage" (PHASE)
37380 Nouzilly France
Tel 33 2 47 42 79 38 (ou 77 96)
Fax 33 2 47 42 79 95
phase@tours.inra.fr
<http://www.inra.fr/internet/Departements/phase/>

"animal"
the new International Journal of Animal Biosciences
<http://www.animal-journal.eu/>
<http://journals.cambridge.org/ANM>