Triphase : co-construction d'une ressource termino-ontologique

Agnès Girard, Claire Nédellec

Triphase : *co-construction d'une ressource termino-ontologique* [1]

La construction de *TriPhase* ou «Terminologie pour la Recherche d'Information du Département Phase», a fait l'objet d'un projet initié par le chef du Département scientifique Physiologie Animale et Systèmes d'Élevage (Phase) de l'Inra (Institut National de la Recherche Agronomique) en 2013. Retour sur les différentes étapes de ce projet collaboratif inédit.

Le domaine de recherche du Département Phase est par nature transversal en termes d'objets étudiés, de niveau d'organisation (de la cellule au système d'élevage en passant par l'organisme animal) et de disciplines mobilisées (biologie, physiologie, endocrinologie, neurosciences, comportement…). En 2013, le chef de Département, Benoît Malpaux, pose le problème du classement des publications des chercheurs par thème et sous-thème et l'évolution de ce classement au cours du temps, en l'absence de thésaurus structuré. Il s'agit d'un besoin récurrent, à des fins d'analyse stratégique, d'évaluation ou de pilotage de la recherche. L'indexation manuelle exigerait une quantité de travail humain très importante compte tenu du volume et de la récurrence du besoin. La solution retenue est alors celle d'une indexation sémantique automatique en texte plein des références des publications, titres et résumés. Pour cela les ingrédients nécessaires sont des outils logiciels d'indexation sémantiques et une ressource terminologie-ontologique structurée, laquelle représente avec précision la structuration thématique du Département. Une indexation structurée hiérarchique permet de visualiser l'évolution et la distribution des thèmes au cours du temps à différents niveaux de détail. La ressource étant intégralement à construire dans un délai court pour respecter l'échéance de l'évaluation du Département, il a été décidé d'utiliser une technologie qui avait fait ses preuves dans un projet précédent d'indexation de journal par une ontologie[2] : extraction automatique des termes candidats des références et organisation de ceux-ci en thésaurus à l'aide du logiciel collaboratif, TyDI (Terminology Design Interface)[3]. Les compétences diverses et complémentaires de l'équipe projet sont à la mesure de l'ambition. Il s'agit de compétences en indexation de documents et en construction de thésaurus apportées par le réseau des six documentalistes du Département, des compétences en ingénierie des connaissances et en développement informatique de système d'information, d'indexation sémantique et d'outils d'aide à la conception d'ontologie apportés par cinq personnes de l'équipe de recherche Bibliome de l'Unité Inra MaIAGE (Mathématiques et Informatique Appliquées du Génome à l'Environnement).

CRÉATION D'UNE ARBORESCENCE DE THÉMATIQUES

La première étape du projet a démarré par la construction manuelle du premier niveau de l'arborescence représentant l'ensemble des thématiques de recherche du Département Phase, à partir du document d'orientation à 5 ans. Cette structure a été importée dans l'outil TyDI, ainsi que les ontologies pertinentes pour le domaine, développées par les scientifiques du Département Phase sur les caractères et les phénotypes des animaux d'élevage dans leur environnement : Atol[4] (*Animal Trait Ontology for livestock*) et Eol (*Environnement Ontology for Livestock*).

ENRICHISSEMENT VIA LES PUBLICATIONS DES CHERCHEURS

La deuxième étape a consisté à enrichir la ressource termino-ontologique *TriPhase* par le vocabulaire utilisé dans les publications des chercheurs du Département. Ces publications ont été exportées depuis ProdInra, la base de données institutionnelle de l'Inra[5]. ProdInra archive les publications des chercheurs de l'Inra et le Département Phase y intègre toute sa production. Le vocabulaire associé aux publications et chargé dans TyDI provient de trois origines : les mots-clés des auteurs des publications, les mots-clés choisis par les documentalistes pour l'indexation dans ProdInra et les termes, mots simples et mots composés, extraits automatiquement par l'outil BioYaTeA des titres et résumés (Golik et al., 2013)[6].

Les documentalistes ont structuré les classes sémantiques de *TriPhase*, sachant qu'une classe sémantique est définie par un concept avec un label (le terme favori) et les termes qui lui sont associés, (synonymes, variations typographiques et traductions). Chaque concept est relié à ses parents (concepts plus généraux) et éventuellement à ses enfants (concepts plus spécifiques). Lors de cette étape, des ressources terminologiques extérieures pertinentes pour le domaine comme le MeSH[7], Agrovoc[8] et Cab Thesaurus[9] ont été consultées régulièrement par les documentalistes afin de structurer, justifier ou arbitrer les choix de peuplement.

[1] Une ressource termino-ontologique est une ontologie légère qui vise à répondre à un objectif précis et est orientée par un point de vue. Elle se base sur trois sources de connaissances : l'expertise humaine, les ressources existantes, les documents d'un corpus – Voir le poster de Tissaoui, A. « Typologie de changements et leurs effets sur l'évolution de ressources termino-ontologiques » présenté au 20e Journées francophones d'ingénierie des connaissances (IC 2009), Hammamet (TN). http://ic2009.inria.fr/docs/posters/Tissaoui_Poster_IC2009.pdf

[2] 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. et Le Bail P.-Y.. « ATOL: the multi-species livestock trait ontology» in *Proceedings of The 6th Metadata and Semantics Research Conference (MTSR 2012),* pages 289-300. Springer Verlag *Communications in Computer and Information Science Serie.* Cadiz, Espagne, 28 au 30 novembre 2012. DOI: 10.1007/978-3-642-35233-1_28

[3] Nédellec C., Golik W., Aubin S., Bossy R., « Building Large Lexicalized Ontologies from Text: a Use Case in Indexing Biotechnology Patents », *International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010),* Volume 6317 of the series *Lecture Notes in Computer Science* pp 514-523, Springer Verlag. Lisbon, Portugal, 11th 15th octobre, 2010.

[4] www.atol-ontology.com/index.php/fr/

[5] http://prodinra.inra.fr/?locale=fr

ÉVALUATION DE LA RESSOURCE PAR L'USAGE

Enfin, la dernière étape du projet s'est concentrée sur l'adéquation de la ressource termino-ontologique TriPhase avec le besoin du Département sous la forme d'un travail itératif d'évaluation par l'usage et d'enrichissement de *TriPhase*. Concrètement, la version courante de *TriPhase* est utilisée pour indexer automatiquement le texte plein des publications des chercheurs du Département. Les termes extraits des documents non indexés sont examinés pour être ajoutés prioritairement à *TriPhase*. La pertinence de l'indexation est également évaluée à travers les deux outils destinés au chef de Département et développés par MaIAGE : un moteur de recherche sémantique « Alvis IR-Triphase » (Figure 1) et un outil d'analyse de corpus permettant une représentation graphique « ANStrat » (Figure 2).

UN TRAVAIL COLLABORATIF FÉCOND

Ce projet a été conduit sur une période de 8 mois. Au terme de celui-ci étaient disponibles une première version de la ressource termino-ontologique *Triphase* distribuée publiquement par le portail AgroPortal, l'outil d'analyse stratégique et le moteur de recherche sémantique. Ces résultats sont les fruits d'une collaboration étroite entre les documentalistes, les chercheurs en informatique et une ingénieure en ingénierie des connaissances. Le projet a répondu aux attentes du chef de Département lui permettant la visualisation des thèmes de recherche à différentes échelles au cours du temps. Il est à noter que la période de 5 ans sur laquelle a été menée l'étude n'est pas une durée suffisante pour observer une évolution significative des concepts liés aux thématiques scientifiques du Département. Des améliorations sont encore à apporter à la ressource *TriPhase*. Celle-ci reste en effet à compléter et à homogénéiser, puis à faire valider par les experts des différents domaines. Cette collaboration a en tout cas été enrichissante pour l'ensemble des acteurs du projet.

Du point de vue des chercheurs en informatique, il a apporté des éléments nouveaux dans l'approfondissement des méthodologies d'utilisation et d'évaluation des outils qu'ils développent, en particulier pour le travail collaboratif et à distance.

Du point de vue des documentalistes, les apports de ce travail collectif ont été nombreux : découverte de l'ingénierie de la connaissance, appropriation de nouvelles technologies… Les documentaliste ont en effet acquis de nouvelles compétences dans l'usage des technologies du web sémantique et ont appris à utiliser de nouveaux outils.

Les documentalistes - chacune spécialisée dans un domaine - ont également partagé et élargi la connaissance de leur Département de recherche. Elles ont aujourd'hui une meilleure compréhension des enjeux scientifiques du Département. Elles ont aussi appris à modéliser les concepts et les sous-concepts relatifs au domaine de recherche du Département ; les outils informatiques utilisés produisant des termes extraits des publications, mais pas l'arborescence de la ressource termino-ontologique.

Au niveau humain, ce projet a renforcé le réseau des documentalistes dans une réelle énergie collective et créé une nouvelle dynamique de collaboration avec l'équipe de recherche en analyse de corpus et ontologie. Il a aussi permis de nombreux échanges avec les chercheurs du Département sur les différentes thématiques de recherche.

Enfin, cette expérience a renforcé notre conviction que les professionnels de l'IST ont un rôle à jouer dans la formalisation des démarches et des connaissances, ceci dans la continuité des savoir-faire documentaires comme la sélection, la structuration, la qualification ou encore la normalisation de l'information.

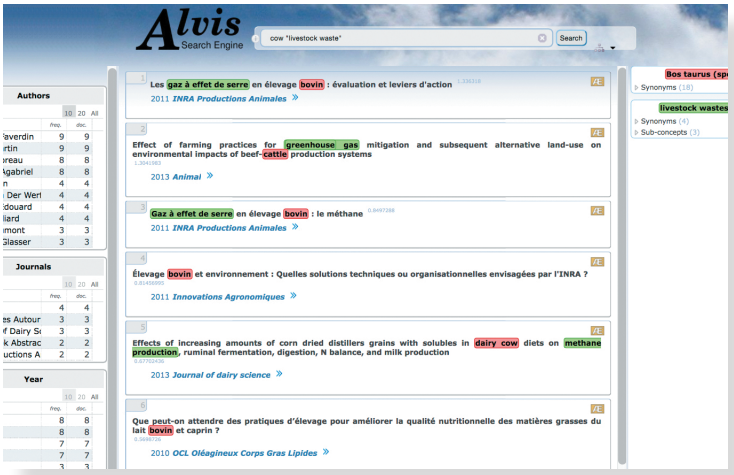

FIGURE 1. Exemple de requête sur moteur de recherche AlvisIR TriPhase.

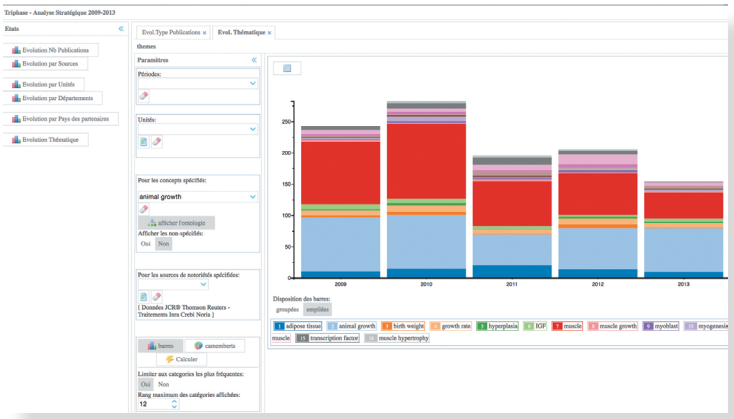

FIGURE 2. Exemple de l'évolution du thème « croissance » au cours du temps dans AnStrat.

AGNÈS GIRARD
Pour le réseau des documentalistes du Département Phase, Inra Rennes
agirard@rennes.inra.fr

CLAIRE NÉDELLEC
Pour l'équipe de recherche Bibliome de l'unité MaIAGE, Inra-Université Paris-Saclay, Jouy-en-Josas
claire.nedellec@jouy.inra.fr

[6] Golik W., Bossy R., Ratkovic Z., Nédellec C. «Improving term extraction with linguistic analysis in the biomedical domain» in *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13),* Special Issue of the journal *Research in Computing Science,* vol 70 ISSN 1870-4069, http://rcs.cic.ipn.mx/2013_70/RCS_70_2013.pdf, 24-30 mars, Samos, Grèce, 2013.

[7] http://www.ncbi.nlm.nih.gov/mesh

[8] http://aims.fao.org/standards/agrovoc/about

[9] http://www.cabi.org/cabthesaurus/