



HAL
open science

Des technologies sémantiques pour l'information scientifique et technique

Claire Nédellec, Agnès Girard

► **To cite this version:**

Claire Nédellec, Agnès Girard. Des technologies sémantiques pour l'information scientifique et technique. Frédoc 2013, Oct 2013, Aussois, France. hal-02805803

HAL Id: hal-02805803

<https://hal.inrae.fr/hal-02805803v1>

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Des technologies sémantiques pour l'information scientifique et technique



Claire Nédellec – Agnès Girard

Aussois – 9 octobre 2013



Des technologies sémantiques pour l'information scientifique et technique

Principes

- Analyse sémantique de texte : un exemple d'utilisation en recherche documentaire
- Une ressource clef : la termino-ontologie
- Acquisition de la termino-ontologie à partir de documents

Le projet TriPhase

- Objectifs d'analyse documentaire en *physiologie animale et élevage*
- Acquisition de termino-ontologie spécialisée, mise en œuvre à l'INRA
- Bilan à trois mois

Conclusion



Indexation *sémantique*

Associe aux documents une carte d'identité *thématique*,
pour la recherche documentaire ou pour des analyses quantitatives.

Alvis *Texte plein indexé par les termes d'une terminologie*

Google Texte plein indexé par des mots

ProdINRA Texte plein indexé par des mots et documents indexés manuellement par des mots-clefs

PubMed Texte plein indexé par des mots et documents indexés manuellement par des mots-clefs structurés en thésaurus (*MeSH*)



Exemple en physiologie animale,

Requête : *dairy product*

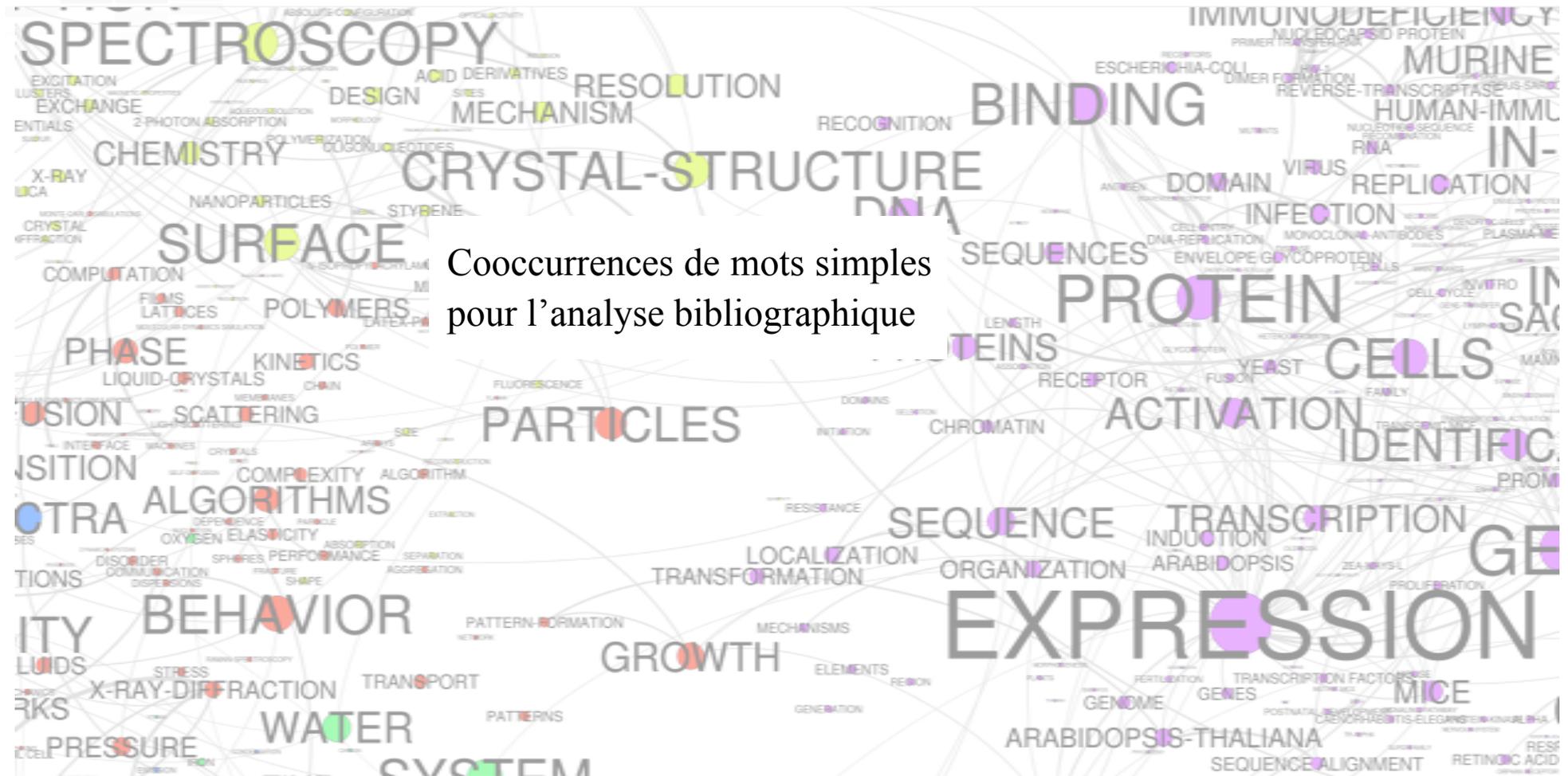
Google **22 documents** : *dairy* et *product* sont des mots simples indexés indépendamment.

... the link between the **product** and the environment. In cheeses, this link is via the feeding of the **dairy** animals ...

ProdINRA **542 documents** (dont 407 mots-clefs) : *dairy* et *product* sont des mots indexés indépendamment dans le texte et *dairy product* est un mot-clef référencé.

TriPhase **757 documents**: *dairy product* est un terme référencé et défini comme *animal product*. Il est indexé dans le corps du texte.

Cows' feeding and **milk** and **dairy product** sensory properties: a review
Cows' feeding and **milk** and **dairy product** sensory properties: a revie...
[Show Categories...](#)



Cooccurrences de mots simples pour l'analyse bibliographique

Copyright Sébastien Grauwil

Query terms : **dairy product** (animal product) [more details...](#)

1-10 among **757** results.

1. **Impact of duration of [milk](#) storage in the mammary gland on [milk](#) composition throughout [milking](#)**
 Impact of duration of [milk](#) storage in the mammary gland on [milk](#) composition throughout [milking](#) [Hide Categories...](#)

Authors	Subject Categories	Concepts	Sources	Years
Dutreuil, M Cebo, C Guinard-Flament, J Hurtaud, C	Agriculture Food Science & Technology	milk(3) mammary gland(1)	WoK	2010
			Journals	
			J DAIRY SCI	

2. **Cows' feeding and [milk](#) and [dairy product](#) sensory properties: a review**
 Cows' feeding and [milk](#) and [dairy product](#) sensory properties: a review...
[Hide Categories...](#)

Authors	Concepts	INRA Units	Years	Sources
Bruno Martin Isabelle Verdier-Metz Anne Ferlay Benoit Graulet Agnes Cornu Yves Chilliard Jean Baptiste Coulon	milk(1) dairy product(1)	URH LRF	2011	prodINRA
			Journals	
			EAAP Book of Abstracts	

3. **Protein supply, glucose kinetics and [milk](#) yield in [dairy cows](#)**
 Protein supply, glucose kinetics and [milk](#) yield in [dairy cows](#) [Show Categories...](#)

Refinement shortcuts

Concepts

Mammary gland morphology
 electronic microscopy
 oxygen level

Species

Anas platyrhynchos
 Catlins group
 Andes

Authors

Brinkmeyer-Langford, CL
 Brown, WC
 Brownstein, MJ

Journals

Journal of Cellular Biochemistry
 CHROMOSOME RES
 J CELL BIOCHEM

Subject Categories

Mathematics
 Genetics & Heredity
 Oncology

Sources

prodINRA
 WoK

Years

2012
 2013
 2008

INRA Units

NuReLiCe
 ERRC



Le terme *Dairy Product* indexe les termes synonymes ou plus spécifiques :

Query

+ **dairy product** (animal product)

Terms details

dairy product

Canonical name: dairy product

Path: /PHASE/TriPhase/animal product/dairy product

Indexes: term

Synonyms

dairy product
milk product
produits laitiers

Sub concepts list

milk casein micelle structure trait
milk fat droplet structure trait
milk fat globule zeta potential
milk rancid flavor intensity
milk bitter flavor intensity
milk astringent flavor intensity
colostrum
immunological importance of colostrum
colostrum immunoglobulin A concentration
colostrum immunoglobulin G concentration
colostrum immunoglobulin M concentration
colostrum yield
colostrum somatic cell count
milk earthy flavor intensity
milk oily flavor intensity
milk malt flavor intensity
milk medicinal flavor intensity
milk structure trait
milk metallic flavor intensity
milk dusty flavor intensity
milk green flavor intensity
milk grainy flavor intensity
milk light-oxidized flavor intensity
milk lack of freshness flavor intensity
milk sour flavor intensity
milk sour aromatics flavor intensity
milk vanilla flavor intensity
milk nutritional quality
milk sweet flavor intensity



Termino-ontologie

Ontologie : un graphe où les nœuds sont des concepts et les arcs des relations entre ces concepts

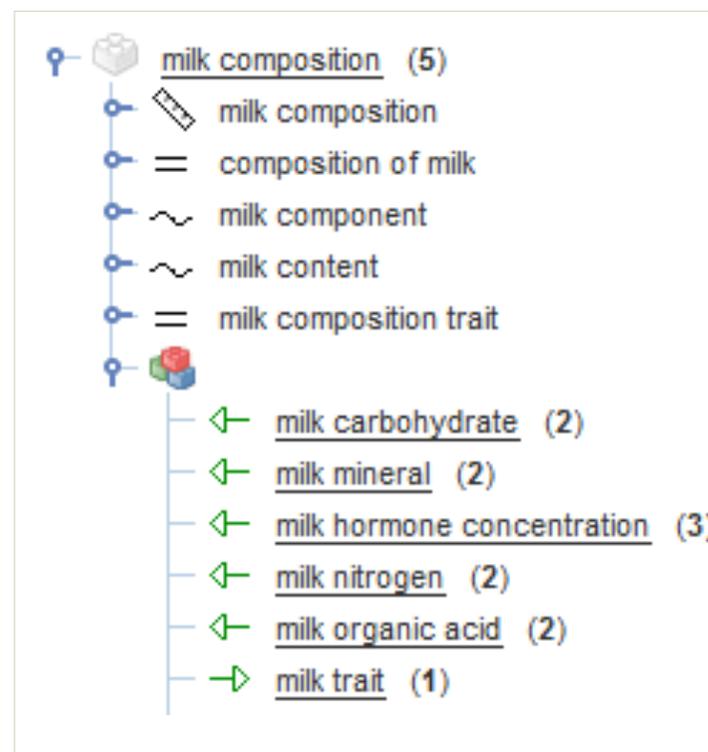
Chaque concept est

Relié à ses parents (concepts plus généraux)

Relié à ses spécialisations (concepts plus spécifiques)

Associé à des informations terminologiques

- Nom du concept (terme en anglais)
- Synonyme exact (=)
- Synonyme approché (~)
- Traduction en français
- Variation typographique





Associer termino-ontologie et documents

L'analyse sémantique identifie les *unités sémantiques* du texte et les associe aux concepts de l'ontologie.

Selection for **Adaptation** to Dietary Shifts: Towards Sustainable **Breeding** of Carnivorous **Fish**

Richard Le Boucher, Mathilde Dupont-Nivet, Marc Vandeputte, Thierry Kerneis, Lionel Goardon, Laurent Labbé, Béatrice Chatain, Marie Josée Bothaire, Laurence Larroquet, Françoise Médale, Edwige Quillet

Abstract

Genetic **adaptation** to **dietary** environments is a key process in the evolution of natural populations and is of great interest in animal **breeding**. In **fish** farming, the use of **fish** meal and **fish** oil has been widely challenged, leading to the rapidly increasing use of plant-based products in **feed**. However, high substitution rates impair **fish** health and **growth** in carnivorous species. We demonstrated that survival rate, mean **body weight** and biomass can be improved in **rainbow trout** (*Oncorhynchus mykiss*) after a single generation of selection for the ability to adapt to a totally **plant-based diet** (15.1%, 35.3% and 54.4%, respectively). Individual variability in the ability to adapt to major **diet** changes can be effectively used to promote **fish welfare** and a more **sustainable aquaculture**.



Conception de termino-ontologie à partir de corpus

- Dans des domaines spécifiques, les termino-ontologies sont rarement suffisantes et complètes.
- Les documents sont une source reconnue de termes et de structuration pour des **méthodes d'acquisition de connaissances** manuelles et automatiques
- Distinguer et relier le niveau lexical, *les termes*, et niveau conceptuel, *l'ontologie*, pour indexer.

Deux étapes

- Extraction automatique de termes (ex. *YaTeA*, *Syntex*)
- Structuration et modélisation (ex. *TyDI*, *Protégé*)



Exemple de TyDI (*Terminology Design Interface*)

TyDI

- Interface collaborative pour valider et structurer des termes et modéliser les concepts
- Pour des experts du domaine en partie autonomes avec l'outil et accompagnés d'un ingénieur de la connaissance.

Principe

- Importer les thésaurus pertinents et importer les termes extraits des documents
- Valider et structurer les termes en classes et en hiérarchies, de façon graphique et collaborative

Démarche réaliste dans un domaine spécialisé

Term grid (1) - free search - Formation ATOL

Semantic Classes of term withdrawal response

Filter

ignore case

Form: include inferred terms: include dismissed terms:

Lemma: include unparsed phrases:

Syntactic category: Word count: >= <=

Head: Nb occurrences: >= <=

Expansion: show only class members: show only class representative:

Prevalidation: Justification: all users:

producer: OBO_1 FastR_2 FastR_1 YaTeA_1 Validation: D D? ? V? V all users:

Semantic Classes Tree Window

Formation ATOL withdrawal

- psychoneuroendocrinological state trait
 - behavior trait
 - biological rhythm trait
 - cognitive functions trait
 - emotional functions trait**
 - metabolism trait
 - pain responses trait
 - antalgic posture
 - emotional functions trait**
 - nociception
 - withdrawal response**

OccurrenceInContext Window - Candidate 8880651

Filename	#	Context
/bibdev/corp...	1 1	05) , while it tended to be correlated with the withdrawal response when approached from the front (P 0 .
/bibdev/corp...	1 1	However , the withdrawal response of the sow when in the farrowing crate was observed by another stockperson .
		This withdrawal response was strongly correlated with the other behavioural responses such as nervousness of sow in the crate before and around farrowing

43 rows

Surface form	Nb occ...	Nb doc.	Head	Expansion	Nb words	Syntactic categ...	ClaireNedellec
'withdrawal crate'	1	1	crate'	'withdrawal	2	JJ NN	
withdrawal movements	1	1	movement	withdrawal	2	NN NNS	
withdrawal period	2	2	period	withdrawal	2	NN NN	
withdrawal reaction	4	3	reaction	withdrawal	2	NN NN	
withdrawal response	5	4	response	withdrawal	2	NN NN	V
withdrawal test	4	1	test	withdrawal	2	NN NN	

Interface TyDI

_02



Le projet TriPhase



Phase en quelques mots

- Département de recherche de l'INRA en *Physiologie animale et systèmes d'élevage*
- Finalité des recherches : conception et l'évaluation de systèmes d'élevage durables (animal, système, socio-économie, environnement)
- 15 unités de recherche + 6 unités sous contrat + des unités expérimentales
- ≈ 350 chercheurs
- ≈ 250 doctorants et post-doctorants accueillis par an
- ≈ 1000 publications par an (45% articles)



TriPhase : contexte et objectif

TriPhase - Terminologie pour la recherche d'information du département Phase

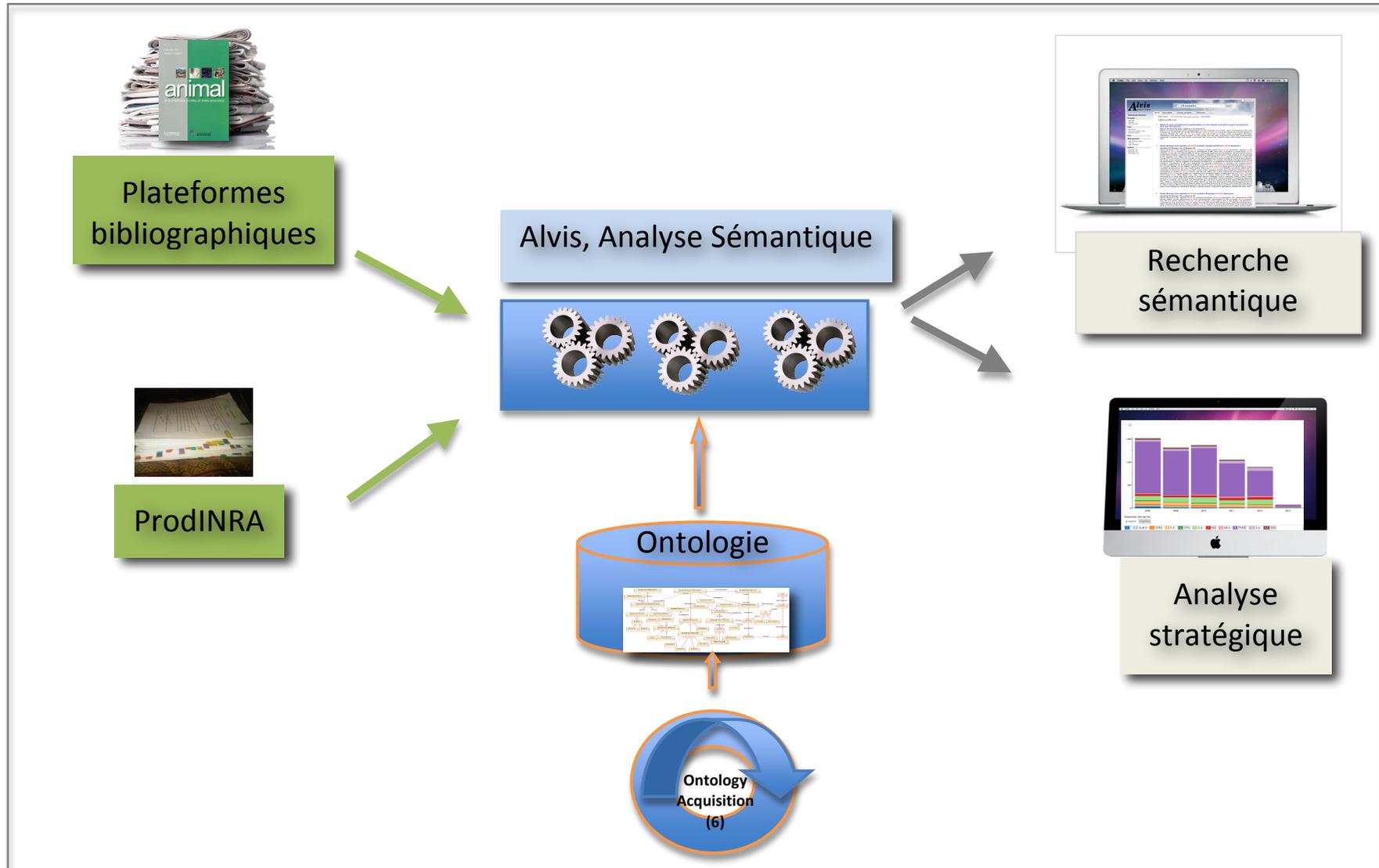
Objectif

- Analyser les publications du département à des fins stratégiques : analyse quantitative des termes au cours du temps, par unité, partenariat, ...
- Disposer d'un moteur de recherche sémantique bibliographique spécialisé

Contexte

- Pas de ressource d'indexation *clef en main*
- 2 ontologies créées par des chercheurs Phase : ATOL *Animal Trait Ontology for Livestock* et EOL *Environment Ontology for Livestock*
- Moteur de recherche ATOL / revue *Animal* ; connaissance des outils développés par MIG
- Un réseau de documentalistes expérimentées dans l'indexation des publications

Architecture TriPhase





Construction de la termino-ontologie *TriPhase*

Implication des documentalistes

- Construire une termino-ontologie : identifier les unités sémantiques dans les publications, leur associer un concept et organiser les concepts dans un modèle.

Corpus

- Collection des publications scientifiques de Phase référencées dans ProdInra et complétées par les publications du WoS (période 2008-2013)

Ressources

- Modèle des thématiques de recherche du département (schéma stratégique 2010-2015) sous forme de carte heuristique.
- Outil collectif d'aide à la construction *TyDI*, après une formation.
- Population de l'ontologie
 - Termes extraits automatiquement à partir du corpus (titres et résumés)
 - Termes d'indexation des notices dans ProdInra par les documentalistes
 - Ontologies *ATOL* et *EOL*
 - Ressources externes : *Mesh*, *Agrovoc*, *Cab Thesaurus* (aide) et *NCBI* (espèces)



Peuplement de l'ontologie par les termes

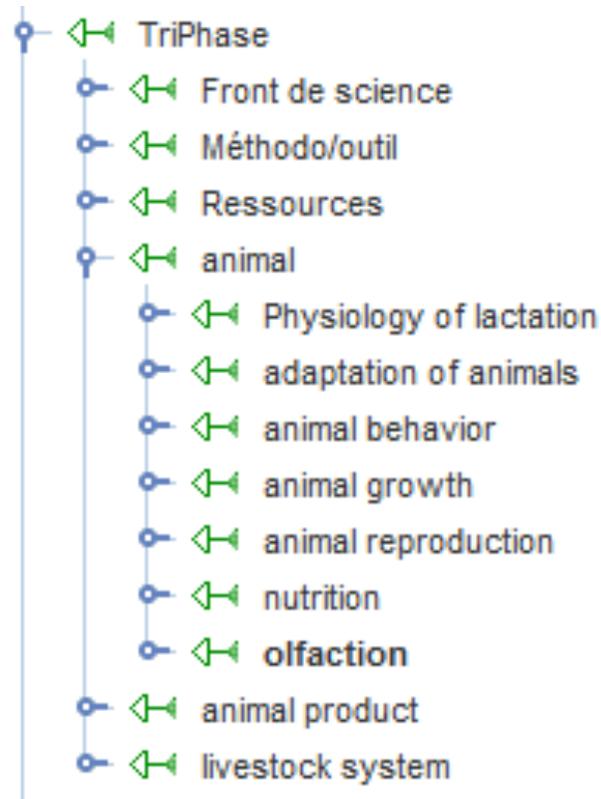
- 1 thématique = 1 documentaliste
- Recherche et regroupement de termes pour chaque thématique (synonyme, reformulation, acronyme, traduction) ainsi que des spécialisations et généralisations des termes du domaine.
- Création et structuration des classes sémantiques : apport des différentes ressources et des scientifiques.
- Mesure de la qualité de la termino-ontologie
 - Le moteur de recherche : identification des incohérences et lacunes,
 - Par projection de l'ontologie sur les documents : termes les plus fréquents des documents non retrouvés, mots clés associés à ces documents, ...

⇒ Travail itératif



L'ontologie *TriPhase*

A ce jour, \approx 1 700 concepts et 2 200 termes



L'ontologie TriPhase au service de l'analyse stratégique

TriPhase - Analyse Stratégique

Etats

- Evolution Nb Publications
- Evolution par Sources
- Evolution par Unités
- Evolution par Départements
- Evolution par Pays des partenaires
- Evolution Thématique

Evol. Thématique x Evol. Thématique x

themes

Paramètres

Périodes:

Au moins un auteur d'une unité spécifiée:

Seulement si premier auteur:

Oui Non

Pour les concepts spécifiés:

animal growth

afficher l'ontologie

Afficher les non-spécifiés:

Oui Non

barres camemberts

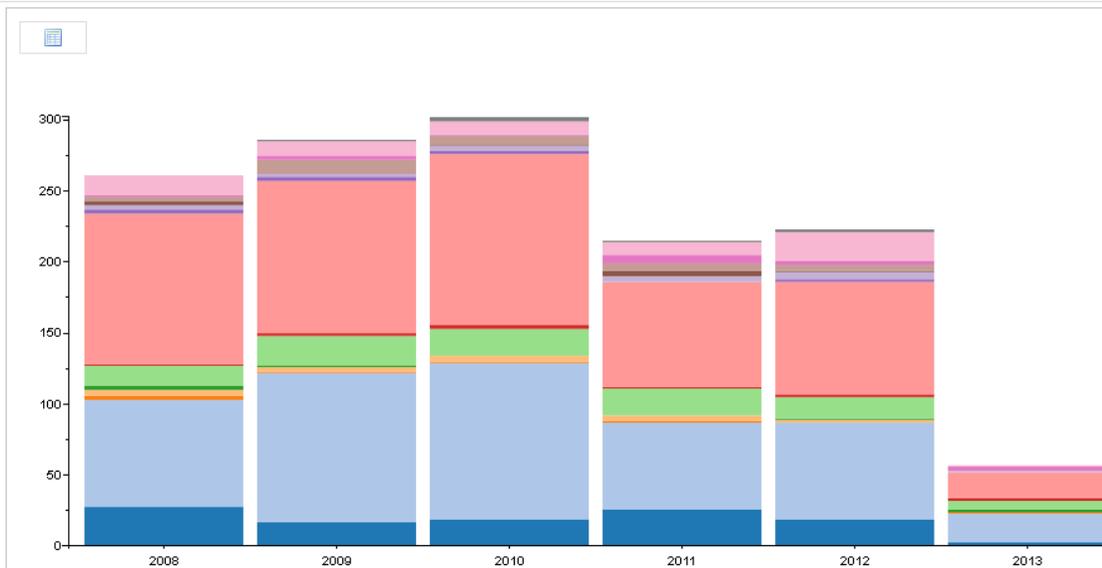
Calculer

Limiter aux catégories les plus fréquentes:

Oui Non

Rang maximum des catégories affichées:

12



Disposition des barres:

groupées empilées

- 1 adipose tissue
- 2 animal growth
- 3 body size
- 4 growth hormone
- 5 growth model
- 6 growth rate
- 7 hyperplasia
- 8 muscle
- 9 muscle atrophy
- 10 muscle growth
- 11 muscle hypertrophy
- 12 myosin
- 13 myostatin
- 14 skeletal muscle
- 15 muscle protein

Interface utilisable par un navigateur (développement MIG)



Ontologie TriPhase, bilan et perspectives

Après deux mois de travail ...

Une termino-ontologie à consolider par un travail de qualité

- Compléter la terminologie : itération entre le moteur de recherche et indexation du corpus par l'ontologie
- Classes sémantiques à retravailler avec des experts du domaine
- Une homogénéisation des classes, traduction des termes
- Une validation par les experts du domaine

La partie logicielle est achevée



Les apports de ce travail collectif pour notre réseau de documentalistes

- Appropriation de nouvelles technologies et découverte de l'ingénierie de la connaissance
- Nouvelles compétences, nouveaux outils
- Apprentissage du travail en complémentarité
- Dynamique de groupe : élément important de notre motivation
- Connaissance des thématiques de *Phase* partagée par toutes
- Dialogue avec les chercheurs sur les thématiques de recherche
- Reste sur des fondamentaux de notre métier : sélectionner, structurer, qualifier, normaliser, ... l'information
- Changement de comportement par rapport à nos pratiques documentaires : Passer des mots-clés pour indexer, aux concepts pour modéliser.



Conclusion (1)

- Les professionnels de l'IST ont un rôle à jouer dans la formalisation des démarches et des connaissances, dans la continuité des savoir-faire documentaires.
- *Ce transfert des compétences sur un terrain nouveau, est plus qu'un simple déplacement. Il implique un changement de paradigme et impose d'acquérir de nouveaux comportements (Dominique Cotte – Documentaliste, 2011).*
- Ce transfert des compétences est possible, en regard de notre expérience avec un collectif très hétérogène (en termes de formation initiale, de parcours,...), accompagné par un ingénieur de la connaissance expérimenté.



Conclusion (2)

- La recherche d'information spécialisée de qualité requiert le développement d'applications spécifiques
- Des méthodes génériques d'indexation sémantique de documents et d'acquisition de connaissances sont opérationnelles.
- Leur utilisation requiert
 - Des **besoins documentaires** clairement identifiés
 - La construction de **ressources spécialisées termino-ontologiques**, validées et mises à jour
 - Des outils assistants, avec des **interfaces homme-machine** appropriées aux besoins



L'équipe au grand complet

Phase, documentalistes	MIG
Pascale Avril Emilie Bernard Maryse Corvaisier Marie-Laure Touzé Nathaële Wacrenier Agnès Girard Et les chercheurs de Phase	Robert Bossy – informaticien Frédéric Papazian – informaticien Wiktorija Golik – ingénieur de la connaissance Claire Nédellec – chercheuse en analyse sémantique



?