Réseau d'écotoxicologie terrestre et aquatique

Fiche thématique N°16 - Aout 2018

TRACE : un Thésaurus pour la Recherche et l'Analyse de Contenus en Écotoxicologie

Introduction

Le réseau ECOTOX développe le projet **TRACE** (**T**hésaurus pour la **R**echerche et l'**A**nalyse de **C**orpus en **E**cotoxicologie), qui porte sur la construction et l'exploitation d'un thésaurus en écotoxicologie.

Un thésaurus est une liste organisée de termes (ou mots-clés) contrôlés représentant les concepts d'un domaine. Les termes sont reliés les uns aux autres par :

- des relations hiérarchiques : terme générique / terme spécifique,
- des relations d'équivalence : terme préférentiel / synonymes - ou descripteur / non descripteur,
- des relations d'association : « voir aussi… ».

Objectifs du projet TRACE

TRACE vise à constituer un référentiel terminologique partagé par le collectif de recherche « Ecotox » et à en faire un élément central du système d'information du réseau Ecotox. Différents usages du thésaurus sont envisagés et testés : aide à la recherche documentaire, indexation et analyse sémantique de documents et de données de recherche, partage de connaissances et aide à la valorisation des productions du collectif auprès de différents utilisateurs (recherche, monde socio-économique…).

Bonnes pratiques : pourquoi utiliser un thésaurus ?

Pour mieux s'informer : suivre et analyser la production scientifique

- Chercher l'information dans des bases de données : comme plusieurs vocabulaires peuvent coexister, il est préférable de combiner dans sa recherche tous les synonymes d'un même concept, et si besoin ajouter des termes génériques ou associés, ou affiner en indiquant des termes plus spécifiques.
  Lors de la construction de la requête, le thésaurus aide à identifier tous les concepts à combiner et leurs synonymes. Il peut aussi permettre d'identifier des équivalents dans une autre langue. Sa structuration sous forme arborescente peut également guider la sélection des termes pertinents.
- Analyser des corpus bibliographiques : la nécessité d'identifier l'information pertinente au sein de corpus toujours plus volumineux a suscité l'émergence de la fouille de texte et de la lexicométrie. Or, l'analyse automatique de corpus est plus efficace si elle s'appuie sur un thésaurus : la prise en compte automatique des synonymes la rend plus robuste. En effet, l'analyse quantitative des mots (répétitions, distributions, associations…) est bien plus pertinente si chaque descripteur est comptabilisé pour chacune de ses formes et chacun de ses synonymes répertoriés dans le thésaurus. Autre avantage, lorsque les outils de recherche (moteur sémantique) et d'analyse le permettent, la prise en compte de la structure hiérarchique du thésaurus permet d'explorer et d'analyser le corpus en navigant dans l'arborescence. Les résultats de l'analyse obtenue sont consolidés : pour chaque terme générique, sont comptabilisées et additionnées les occurrences de chacun de ses termes spécifiques.

L'existence du thésaurus TRACE permet donc d'envisager son utilisation avec des outils de text-mining ou de lexicométrie. Avec une telle ressource, nous sommes à même de produire rapidement des analyses de tendances et des indicateurs venant en appui de travaux de synthèse ou de prises de décision.

Pour mieux communiquer : rédiger avec un vocabulaire précis, reconnu et partagé par sa communauté scientifique, utiliser des synonymes et enrichir ses publications de mots-clés appropriés contribuent à leur notoriété. Mieux indexées par les moteurs de recherche et les bases de données, elles figurent plus souvent dans la liste des résultats.

Le projet TRACE en aout 2018

Il se décline en 5 lots.

LOT 1 : identifier et collecter des vocabulaires en lien avec l'écotoxicologie

Nous avons collecté et exploité différents lexiques, glossaires et bases de données disponibles au niveau international et en lien avec l'écotoxicologie (polluants, milieux, organismes…) tels que « IUPAC Glossary of Terms Used in Toxicology[1] » ou TAXREF[2]. Voir la liste de ressources sélectionnées sur le site du réseau Ecotox.

LOT 2 : constituer le thésaurus

Nous avons constitué le thésaurus en mobilisant les compétences des scientifiques du réseau autour d'une architecture comprenant une dizaine de sous-domaines :

- compartment
- effect
- pollutant
- research
- organism
- pollutant fate
- vulnerability modulating factor
- biological level
- exposure
- ecotoxicology societal application

Chaque sous-domaine est développé sur au moins deux sous-niveaux d'arborescence.

La construction du thésaurus s'est faite par itérations successives. Certains descripteurs proviennent des vocabulaires identifiés dans le Lot 1 et d'autres d'articles scientifiques analysés par les veilleurs du réseau. Chaque sous-domaine a été relu et validé par des scientifiques du réseau. Enfin, nous avons complété le thésaurus en intégrant des parties de référentiels sur les milieux, les matières actives, les organismes...

Le thésaurus, actuellement en anglais, compte plus de 8000 concepts. Il sera enrichi avec des termes en français et si besoin en latin. Le thésaurus a été construit au moyen de l'outil Luxid Webstudio[3]. Une migration sous VocBench[4] est envisagée pour la maintenance du thésaurus. Le logiciel Word est utilisé pour le travail de validation par les experts.

LOT 3 : déployer le thésaurus dans le réseau

Ce lot est destiné mettre à disposition des chercheurs du réseau Ecotox un outil facile à utiliser pour permettre son appropriation et *in fine* son utilisation dans plusieurs domaines :

- Aide à la publication dont ajout de mots-clés aux publications,
- Aide à la recherche d'information par la constitution de requêtes dans les bases de données bibliographiques (WoS, PubMed, etc.)
- Appui pédagogique pour les stagiaires, média de communication.

Deux types de publications en ligne du thésaurus sont envisagées : sous forme de document textuel (pdf) et sous forme informatique pour une consultation interactive.

Le déploiement du thésaurus auprès de la communauté passe nécessairement par des étapes de communication et d'incitation, en particulier au sein du réseau Ecotox.

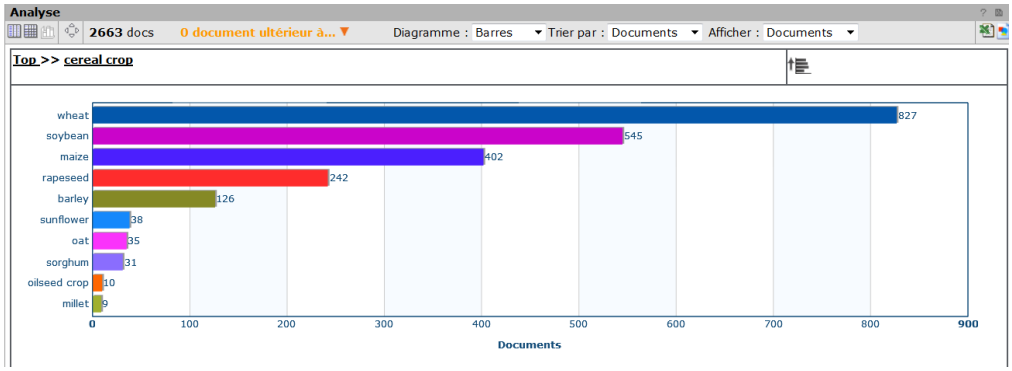
LOT 4 : analyser à l'aide du thésaurus les contenus bibliographiques

Ce lot vise à étudier la faisabilité d'utiliser des outils d'analyse textuelle capables d'exploiter le thésaurus pour analyser des corpus documentaires. Nous cherchons à en tirer une connaissance globale des thématiques traitées dans l'ensemble du corpus, des sous-thématiques et des liens entre elles ainsi que leur évolution dans le temps. Ces outils nous permettent aussi d'identifier des groupes de documents ou « clusters » partageant les mêmes thématiques (traditionnelles ou émergentes). Les analyses produites sont à destination des chercheurs et des managers (responsables d'équipes, directeurs d'unité, département, directions scientifiques…) afin de leur fournir une aide au pilotage ou à l'évaluation (rédaction de synthèses bibliographiques et dossier d'évaluation de collectifs par exemple).

Un exemple de mise en pratique

Nous avons utilisé un logiciel de fouille de texte (tel que Luxid Information Analytics ou Temis) que nous avons alimenté avec le thésaurus ECOTOX et fait travailler sur un corpus de 2263 notices issues du Web of Science portant sur les **semences enrobées.** Voici quelques exemples des résultats obtenus.

Répartition des 2263 publications sur les semences enrobées par culture

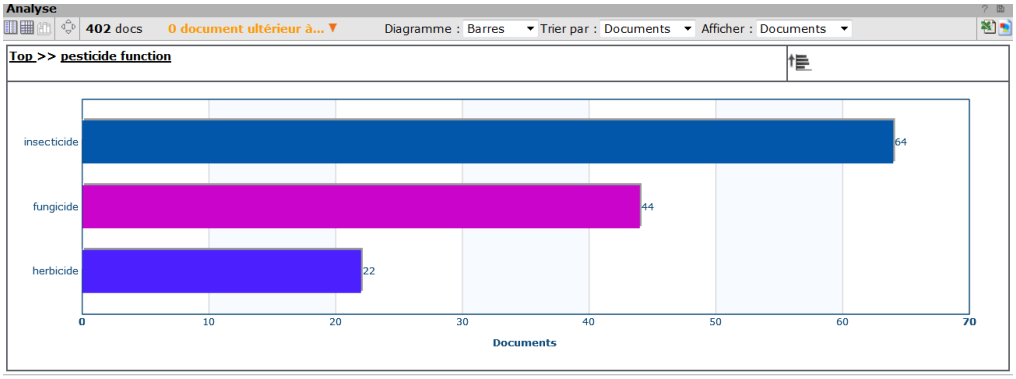


Réseau d'écotoxicologie terrestre et aquatique

Fiche thématique N°16 - Aout 2018

1-Travail centré sur une culture

-Pour le maïs enrobé (402 ref), quelles sont les catégories des pesticides citées ?



Analyse sur le concept pesticide function (dont les sous-thèmes sont *acaricide, fungicide, herbicide, insecticide, larvicide, molluscicide, et rodenticide* dans le thésaurus). Seules trois catégories sont citées dans les notices.

-Pour le maïs enrobé (402 ref), quelles relations *active substance / individual effect ?*

Analyse — 402 docs — Nb docs min. : 1 — Colonne : biomarker at individual level — Pourcentage — OK

Filtrer — biomarker at individual level — active substance

active substance	behavior	bio chemical marker	growth	Life history trait	physiological change
2,4-D			1		
ascorbic acid			1		
azoxystrobin	1		2	1	
bifenthrin				1	
blood meal		1			1
captan		1	4		
carbendazim			1		
carbofuran	1		1	3	
carboxin			1		
chlorantraniliprole	1			2	

-Pour le maïs enrobé (402 ref), quelles relations *plant component / active substance ?*

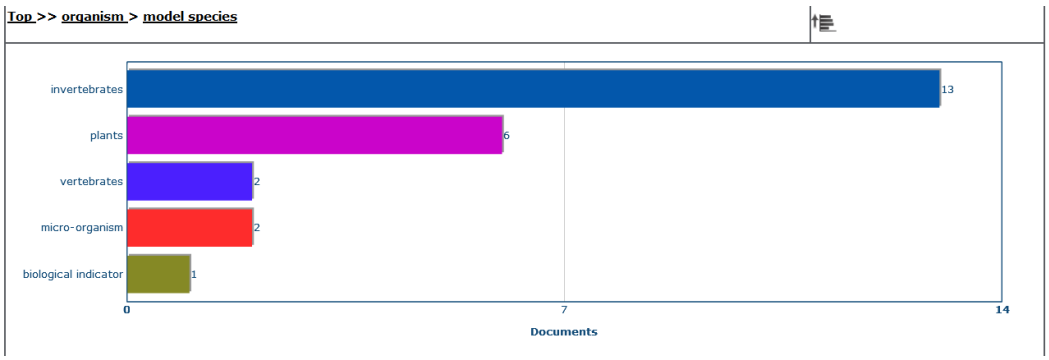
Liste de documents — Analyse — Analyse Temporelle — Clustering — Proximité

Analyse — 402 docs — Nb docs min. : 1 — Colonne : plant component — Pourcentage — OK

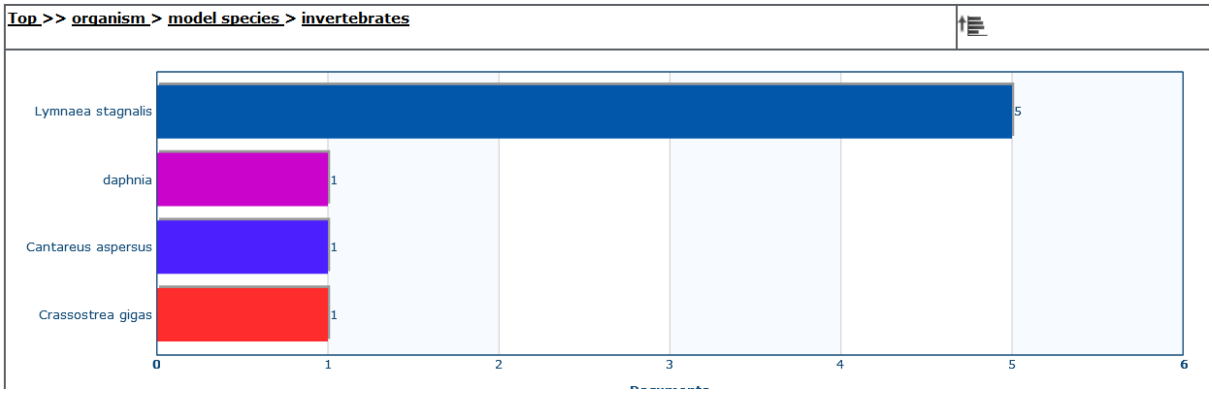
active substance	exudate	leaf	plant tissue	root	seed	shoot
sodium hypochlorite						
sunflower oil					0.9 %	
tebuconazole					1.8 %	
thiabendazole					0.9 %	
thiacloprid					0.9 %	
thiamethoxam	28.6 %	19.0 %	33.3 %	7.4 %	7.3 %	
thiophanate-methyl					0.9 %	
thiram				3.7 %	5.5 %	
tolclofos-methyl				3.7 %		
triadimenol					0.9 %	
tribenuron					0.9 %	
trifloxystrobin					0.9 %	
trinexapac		4.8 %	33.3 %	11.1 %	10.1 %	
triticonazole				3.7 %	1.8 %	
urea		19.0 %		7.4 %	4.6 %	14.3 %

2-Autre approche : analyse sur le sous corpus semences enrobées ET les effets non-cibles

-Répartition des effets non-cible par type d'organismes



-Focus sur les invertébrés : 8 ont été identifiés, mais la liste des invertébrés doit être complétée. La qualité de l'analyse dépend de l'exhaustivité du thésaurus.



Remarques sur ces analyses

Une fois le corpus chargé dans le logiciel de fouille de texte et annoté avec le thésaurus, ces analyses sont disponibles en un clic.

A chaque étape, on peut accéder aux notices, lire les résumés.

En naviguant dans l'arborescence du thésaurus et en produisant différents types d'analyse, on peut explorer le corpus pour avoir une vision quantitative solide des thèmes traités et répondre à des questions précises :

- quelles sont les cultures les plus étudiées, avec quelles substances actives sont-elles traitées, avec quels effets ?
- Pour le maïs, quels sont les effets des enrobages sur la plante, sur quelles parties de la plante ?
- Pour les effets non-cible, quels sont les organismes impactés, et parmi les invertébrés, lesquels sont le plus étudiés ?

Importance de la qualité du thésaurus

Plus le thésaurus est riche, détaillé, avec des synonymes, plus les résultats sont pertinents.

WP5 : intégrer ces ressources dans le dispositif de veille

L'objectif du Lot 5 est d'utiliser le thésaurus pour optimiser le dispositif de collecte (filtres/agents/indexation automatique), le plan de classement et les livrables (tags). De manière comparable à l'expérimentation sur les notices bibliographiques, un outil de fouille de texte associé au thésaurus, permettrait aussi d'analyser les résultats de la veille et d'en réaliser des synthèses ou d'en extraire des indicateurs exprimant, par exemple, l'évolution des thématiques du domaine.

Conclusion

Nous vous avons présenté l'intérêt que présente un thésaurus pour le chercheur, notamment par les potentialités qu'il offre, associé à des outils informatiques, pour analyser les contenus de grands corpus bibliographiques.

Nous souhaitons que ce projet, collaboratif, s'appuie maintenant sur la communauté des chercheurs en écotoxicologie pour prendre de l'ampleur.

Contacts

Christine Sireyjol[1], Sophie Aubin[2], Christian Mougin[1]

[1]UMR ECOSYS, INRA, AgroParisTech, Université Paris-Saclay, 78026, Versailles, France

[2]DIST, INRA, 49070, Beaucouzé, France

Pour en savoir plus

[1]https://sis.nlm.nih.gov/enviro/iupacglossary/glossarya.html

[2]https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref

[3]https://www.expertsystem.com/resource/brochure-luxid-webstudio/

[4]http://vocbench.uniroma2.it/