Politique d'indexation avec le thésaurus INRAE dans HAL INRAE

Sonia Bravo, Véronique Decognet, Olivier Dupré, Agnès Girard, Roselyne Tâche, Clotilde Nicol

Politique d'indexation avec le thésaurus INRAE dans HAL INRAE

Guide à l'usage des professionnels de l'information scientifique et technique

I. Introduction

Ce guide a été élaboré en 2021 par un groupe de travail constitué de six professionnels du réseau IST@INRAE : Sonia Bravo, Véronique Decognet, Olivier Dupré, Agnès Girard, Clotilde Nicol et Roselyne Tâche. Ce document est évolutif. Pour toute suggestion, vous pouvez contacter mailto:clotilde.nicol@inrae.fr.

Il a pour objectif d'apporter un cadre d'indexation aux professionnels IST qui modèrent et patrouillent les dépôts dans HAL INRAE afin d'harmoniser les pratiques sur l'usage des champs mots-clés et du thésaurus INRAE, plus particulièrement sur leur articulation et complémentarité. Il s'appuie sur le document Utilisation du thésaurus INRAE et autres champs d'indexation du portail et le complète avec des cas d'usage précis.

Dans les limites de ce périmètre, il n'a pas vocation à expliquer le processus d'indexation d'un document et n'apporte pas de recommandations sur l'indexation des autres champs de HAL INRAE : domaines de HAL, mots-clés Mesh, classifications.

II. Thésaurus INRAE

Quelques définitions utiles

Mot-clé [1] **:** mot ou expression choisi généralement dans le titre ou le texte d'un document pour en caractériser le contenu et en permettre la recherche. Il fait partie du langage naturel, libre et est à distinguer d'un descripteur, qui est un terme normalisé dans un thésaurus.

Descripteur [1] **:** terme retenu dans un thésaurus pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire. Ce peut être un nom commun ou un nom propre (nom géographique, de société, de personne, terme taxonomique...), une locution, un mot composé ou un groupe de mots.

Concept [2] : exprimé par un terme préféré ou descripteur et des termes non-descripteurs. Il a des relations hiérarchiques et associatives avec d'autres concepts. Il est aligné avec les concepts d'autres thésaurus par des relations d'équivalences. Il peut être défini et documenté par différents types de notes.

Présentation du thésaurus INRAE

Le thésaurus INRAE propose un système d'indexation précis et structuré couvrant les différents domaines de recherche INRAE. Il est composé de plus de 15 000 concepts, tous identifiés de manière unique et pérenne par un Uniform Resource Identifier (URI). Chaque concept, qui correspond à une notion/idée précise unique, peut être décrit par plusieurs termes et dans plusieurs langues (français, anglais) : un terme préférentiel (un par langue) qui s'affichera par défaut et des termes alternatifs (synonymes, acronymes).

[1] Définitions de l'Association des professionnels de l'information et de la documentation

[2] https://opentheso.hypotheses.org/67#langages-documentaires

Exemples de concepts :

Terme préférentiel (fr)	Synonyme(s)	Terme préférentiel (en)
agroécosystème	agrosystème écosystème agricole	agroecosystem
nutrition animale	alimentation animale	animal nutrition

HAL INRAE intègre dans tous les formulaires de dépôt un champ spécifique « Thésaurus INRAE », visible uniquement sur le portail de l'Institut, mais interrogeable *via* la recherche simple de tout portail HAL.

Pourquoi indexer avec le thésaurus INRAE ?

De par son essence, le thésaurus INRAE est le référentiel approprié pour décrire de manière pertinente les productions scientifiques de l'Institut. Grâce à sa division thématique, il pourra aider un indexeur à mieux appréhender un domaine et son environnement sémantique lorsque celui-ci lui est peu familier. Il permet ainsi de trouver les termes qui conviendront le mieux à l'indexation, notamment en s'appuyant sur les définitions des concepts, visibles depuis le portail de consultation du thésaurus.

L'indexation avec le thésaurus INRAE va faciliter et améliorer la performance de la recherche bibliographique :

- la recherche sur des descripteurs du thésaurus interroge, en effet, à la fois les termes préférentiels mais aussi les synonymes et traductions. Par exemple, pour le concept « nutrition animale », il est possible de retrouver toutes les publications indexées avec ce concept aussi bien en cherchant « nutrition animale » qu'à « alimentation animale » ou « animal nutrition » ;
- les termes sont issus d'un vocabulaire contrôlé et non d'un vocabulaire libre comme pour le champ mots-clés (mots-clés des auteurs ou proposés par l'indexeur). L'indexation en vocabulaire libre est aisée car elle utilise la langue naturelle, avec son lexique très étendu, ses variations grammaticales, sémantiques… Mais, cette richesse de la langue naturelle risque d'induire une perte d'efficacité lors d'une recherche thématique (moins bonne exhaustivité des références, références non pertinentes) à moins de mettre en place une stratégie de recherche complexe ;
- un autre argument est lié à la recherche simple dans HAL qui interroge le champ thésaurus INRAE au même titre que les champs titre, résumé… et ce quel que soit le portail HAL consulté ;
- dans le portail HAL INRAE, il est aussi possible de faire une recherche avancée sur le seul champ thésaurus INRAE, ce qui permet d'affiner la liste des résultats sur les concepts interrogés ;
- enfin, les concepts du thésaurus étant déclinés en langue anglaise, le thésaurus INRAE peut servir de dictionnaire anglais-français pour construire des requêtes de recherche dans HAL ou autre base de données ; il contribue ainsi à améliorer la recherche pour des scientifiques non francophones.

En conclusion, le groupe de travail recommande très fortement l'indexation avec le thésaurus INRAE. Celle-ci ne se substitue pas à l'indexation avec les mots-clés libres mais vient la complémenter. En effet, il est nécessaire d'enrichir le champ mots-clés car celui-ci est visible dans tous les portails HAL, à l'inverse du champ thésaurus INRAE visible uniquement dans le portail HAL INRAE.

III. Qu'est-ce qu'une politique d'indexation ?

L'Association française de normalisation (AFNOR) définit l'indexation comme *"l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts évoqués dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse" (1993)*.

La politique d'indexation est l'ensemble des recommandations, des directives précises et des règles imposées aux indexeurs dans une collection particulière, un milieu particulier, une institution particulière… Le groupe de travail s'est appuyé sur la définition proposée par Hudon (1997-1998) [3] pour élaborer la politique d'indexation avec le thésaurus INRAE.

"La politique d'indexation se présente sous la forme d'un document ou d'un ensemble de documents plus ou moins formels contenant les réponses aux questions habituelles que l'indexeur peut se poser au moment de prendre des décisions quant au type, au nombre et à la nature des concepts à retenir pour indexation.
La politique d'indexation est généralement d'application locale et elle inclut normalement des éléments d'information sur l'environnement (types de documents, besoins des utilisateurs, aspects informatiques, etc.), sur les objectifs du travail d'indexation, sur les aides à l'indexation (grille d'indexation, bordereau de travail, langage documentaire).
La politique d'indexation fournit des directives précises sur les sources et les méthodes à privilégier pour l'analyse du contenu, sur les niveaux de profondeur et de spécificité de l'indexation, sur les sources et les méthodes de traduction en langage d'indexation."

D'un abord très pratique, cette définition mentionne les contours que doit prendre le document de politique d'indexation, l'application au contexte local (portail HAL INRAE), les spécificités propres de l'indexation (thésaurus INRAE), les aides aux indexeurs… Elle met l'accent sur le questionnement de l'indexeur quant aux choix possibles d'indexation par rapport à l'outil et au langage documentaire à sa disposition.

[3] Hudon, M. (1997-1998). Indexation et langages documentaires dans les milieux archivistiques à l'ère des nouvelles technologies de l'information. Archives, 29, 75-98
http://www.archivistes.qc.ca/revuearchives/vol29_1/29-1-hudon.pdf

IV. Politique d'indexation avec le thésaurus INRAE

Périmètre

Différents champs dans HAL INRAE permettent de décrire une ressource :

- domaines dans HAL (description des disciplines) ;
- mots-clés (mots-clés auteur issus de la publication ou renseignés par l'indexeur) ;
- thésaurus INRAE ;
- mots-clés MeSH (Medical Subject Headings : thésaurus de référence dans le domaine biomédical) ;
- classifications PACS (Physics and Astronomy Classification Scheme) et MSC (Mathematics Subject Classification).

Le présent document a pour objectif d'aider les professionnels à harmoniser leurs pratiques d'indexation avec les champs mots-clés et thésaurus INRAE, à l'aide de cas d'usages. Pour les autres champs, l'indexeur pourra s'appuyer sur le document Utilisation du thésaurus INRAE et autres champs d'indexation du portail

Cas d'usages

Les cas d'usages ont été définis par le groupe de travail après analyse d'un corpus d'une centaine de dépôts avec ou sans fichiers, indexés dans HAL INRAE avec le thésaurus INRAE (Annexe 1). Ce travail a permis de mieux appréhender (1) l'appropriation du thésaurus INRAE par les indexeurs depuis son intégration dans HAL INRAE, (2) l'articulation entre les champs mots-clés et thésaurus INRAE, (3) le nombre de mots-clés et de descripteurs utilisés.

Cas d'usage 1 - Mots-clés (MC) auteurs présents dans la publication ou renseignés par les auteurs lors du dépôt

- Champ libre mots-clés :
  - copier-coller les MC auteurs de la publication;
- Champ thésaurus INRAE :
  - dupliquer les MC auteurs : rechercher les MC dans le thésaurus INRAE et les sélectionner s'ils sont présents ; à défaut, rechercher un concept synonyme ou voisin
  - enrichir l'indexation avec des concepts du thésaurus INRAE en recherchant soit des termes plus génériques, soit des termes plus précis à partir du titre, résumé, MC Mesh…

Exemples d'application

Cas d'un article de revue indexé avec des MC auteurs

Knock out of specific maternal vitellogenins in zebrafish (Danio rerio) evokes vital changes in egg proteomic profiles that resemble the phenotype of poor quality eggs https://hal.inrae.fr/inserm-03218476 (notice consultée le 27/07/2021)

- Champ libre mots-clés : MC auteur de la publication *Zebrafish, Vitellogenin, Knock-out, CRISPR/Cas9, Proteomics, LC-MS/MS, Egg quality*
- Champ thésaurus INRAE :
  - des concepts voisins des MC auteurs sont recherchés dans le thésaurus INRAE

MC auteurs	Concepts du thésaurus INRAE
Zebrafish	Danio rerio (fr) - Danio rerio (en)

Vitellogenin	vitellogénine (fr) - vitellogenin (en)
Knock-out	knock-out (fr)
CRISPR/Cas9	CRISPR-Cas9 (fr)
Proteomics	protéomique (fr) - proteomics (en)
Egg quality	qualité des œufs (fr)
LC-MS/MS	chromatographie liquide couplée à la spectrométrie de masse en tandem (fr) - liquid chromatography coupled to tandem mass spectrometry (en) syn. LC-MS/MS (fr)

- le champ est enrichi avec des concepts du thésaurus INRAE plus génériques :
  - développement biologique (fr) - biological development (en)
  - embryogenèse (fr) - embryonic development (en)
  - Pisces (fr) - Pisces (en) syn. poisson (fr) - fish (en)

<u>Cas d'une communication à un congrès indéxée avec des MC libres lors du dépôt par les auteurs</u>

Hippo pathway-mediated regulation of micropyle formation by microRNA 202 (miR-202) in the fish oocyte https://hal.inrae.fr/hal-03248173v1 (notice consultée le 27/07/2021)

- Champ libre mots-clés : *micropyle, miR-202*
- Champ thésaurus INRAE : les deux termes *micropyle, miR-202* ne sont pas présents dans le thésaurus INRAE. Le champ est enrichi avec des concepts plus génériques :
  - du terme *miR-202* : *ARN non codant*
  - extraits du titre ou du document : *Pisces, Danio rerio, ovocyte*

Cas d'usage 2 - Les mots-clés (MC) auteurs ne sont pas présents dans la notice ou dans le document (par exemple un rapport, une page Web…)

- Extraire des MC d'après le titre et le résumé s'il est renseigné dans la notice.
- Champ libre mots-clés : saisir ces MC.
- Champ thésaurus INRAE :
  - rechercher ces MC dans le thésaurus INRAE et les sélectionner s'ils sont présents ; à défaut, rechercher un terme synonyme ou voisin.
  - enrichir l'indexation avec les concepts du thésaurus INRAE en recherchant soit des termes plus génériques, soit des termes plus précis

Exemple d'application

Cas d'une communication à un congrès sans document ni résumé

A TILLING approach to generate broad-spectrum resistance to potyviruses in tomato is hampered by eIF4E gene redundancy https://hal.inrae.fr/hal-03195976v1 (notice consultée le 27/07/2021)

- Extraire des MC d'après le titre : Tilling, Potyvirus, Tomato, eIF4E, genetic resistance.
- Champ libre mots-clés : saisir ces MC.
- Champ thésaurus INRAE :
  - rechercher ces MC dans le thésaurus INRAE et les sélectionner s'ils sont présents ; à défaut, rechercher un terme synonyme ou voisin.

MC	Concepts thésaurus INRAE
Tilling	tilling (fr) - targeting induced local lesion in genomes (en)
Potyvirus	Potyvirus (fr) - Potyvirus (en)
Tomato	tomate (fr) - tomato (en)
eIF4E	interaction eIF4E-VPg (fr)
genetic resistance	résistance génétique (fr)

  - enrichir l'indexation avec les concepts du thésaurus INRAE en recherchant soit des termes plus génériques, soit des termes plus précis :
    - culture légumière (fr)
    - virus phytopathogène (fr) - phytopathogenic virus (en)

Mots-clés	en Tilling, Potyvirus, Tomatoe, eIF4E-VPg, genetic resistance
Thésaurus Inrae	• **Potyvirus** (fr) - **Potyvirus** (en) *syn.* virus y pomme de terre (fr) • **interaction eIF4E-VPg** (fr) • **tomate** (fr) - **tomatoe** (en) • **culture légumière** (fr) • **virus phytopathogène** (fr) - **phytopathogenic virus** (en) • **tilling** (fr) - **targeting induced local lesion in genomes** (en) • **génétique de la résistance** (fr)

Cas d'usage 3 - Cas particuliers des mots-clés géographiques, temporels, historiques ou des noms de personnes

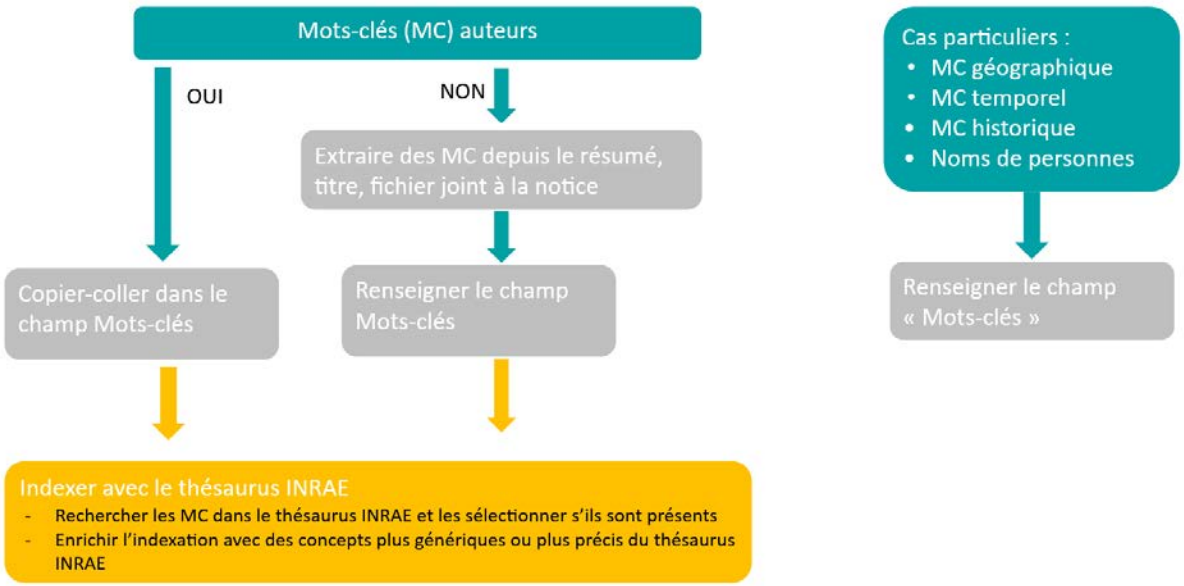
Le thésaurus INRAE n'intègre pas ces descripteurs. Aussi, si l'indexeur souhaite enrichir la notice avec ce type de descripteur, il devra les renseigner dans le champ mots-clés.

Exemples d'application

- Descripteur géographique : Brazil, Cameroon, French Guiana https://hal.inrae.fr/ird-01143899v1
- Descripteur historique : XIXème siècle https://hal.inrae.fr/hal-03147904v1
- Nom de personne : Etienne-Jean Delécluze https://hal.inrae.fr/hal-01989944

Arbre de décision

L'arbre de décision synthétise les cas d'usage décrits ci-dessus.



V. Recommandations

Profondeur de l'indexation

Lors de l'indexation, il est préconisé de renseigner au minimum trois concepts et au maximum vingt concepts (champs mots-clés et thésaurus INRAE confondus).

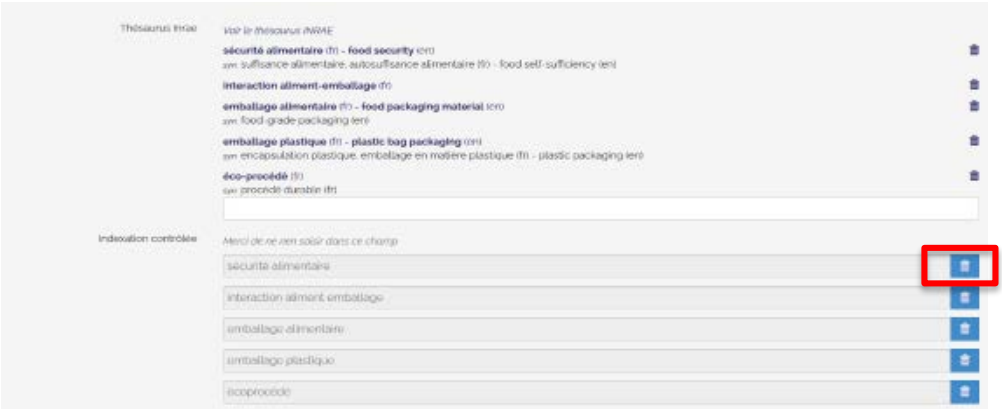
Une suggestion de correction ou de création d'un concept ?

Les utilisateurs du thésaurus INRAE peuvent soumettre des demandes de corrections sur un concept, libellé devenu désuet, traduction incorrecte, ajout d'information sur un concept (traduction absente, synonyme) … Ils peuvent également proposer la création d'un nouveau concept. Les remarques et souhaits d'évolution peuvent être adressés à l'équipe THESAURUS-INRAE *via* le formulaire https://consultation.vocabulaires-ouverts.inrae.fr/thesaurus-inrae/fr/feedback

Champ Indexation contrôlée

Dans HAL INRAE, le champ Indexation contrôlée a été créé lors de la migration des données pour reprendre les concepts des thésaurus Inra et Irstea.
Ex : https://hal.inrae.fr/hal-02617991

Dans le cas où un indexeur souhaite actualiser un dépôt issu de la migration, il est préconisé de rechercher les termes du champ Indexation contrôlée dans le thésaurus INRAE. Si les concepts équivalents sont retrouvés, il alors possible de les supprimer du champs indexation contrôlée.



Qualité de l'indexation : que faire quand un dépôt est réalisé depuis un autre portail ?

Le groupe de travail recommande d'indexer les champs mots-clés et thésaurus INRAE pour toutes les notices INRAE.

ANNEXE 1 DÉMARCHE DU GROUPE DE TRAVAIL

En avril 2021, le groupe de travail a analysé un corpus d'une centaine de dépôts avec ou sans fichiers et indexés dans HAL INRAE selon les possibilités suivantes :

1. Dépôt avec renseignement des domaines de HAL (obligatoire) + champ mots-clés auteur + concepts du thésaurus ;
2. Dépôt avec renseignement des domaines de HAL et du thésaurus, sans utilisation des autres champs d'indexation ;
3. Mixage de ces deux catégories nécessitant l'utilisation de descripteurs géographiques, temporels ou historiques car indexation non couverte par le thésaurus.

Cette analyse montre l'usage et les combinaisons entre les différents champs d'indexation (pluri indexation) et apporte une visibilité sur le nombre de descripteurs utilisés en moyenne, tous champs confondus. Voici les observations obtenues avec cette grille de lecture.

1. Dépôts avec combinaisons des domaines HAL, des champs mots-clés et thésaurus INRAE

- Les mots clés auteur présents dans le document sont copiés/collés dans le champ mots-clés, qu'ils soient en français ou en anglais. Dans un deuxième temps, les indexeurs se sont appuyés sur ces mots-clés pour rechercher les concepts dans le thésaurus. Ex : https://hal.inrae.fr/hal-03191052v1
- Quand les mots-clés de la publication ou les mots clés libres renseignés dans le champ mots-clés sont très spécifiques, le thésaurus permet d'indexer avec des concepts plus génériques. Ex : https://hal.inrae.fr/hal-03191052v1
- Quand les mots-clés du champ mots-clés sont en anglais, leur traduction ou équivalent en français sont recherchés dans le thésaurus. Ex: https://hal.inrae.fr/hal-03162464v1
- Le résumé peut être aussi un point d'entrée pour compléter l'indexation avec le thésaurus INRAE. Ex : https://hal.inrae.fr/hal-03187835v1

2. Dépôt sans document, sans mot-clé et sans résumé, mais indexé avec le thésaurus INRAE

- Dans ce cas l'indexeur s'est appuyé sur le titre pour procéder à l'indexation avec le thésaurus. Ex : https://hal.inrae.fr/hal-03185436v1

3. Dépôts avec combinaisons des domaines, du champ mots-clés géographiques, temporels ou historiques

- Les descripteurs géographiques, temporels ou historiques n'étant pas présents dans le thésaurus, ils sont renseignés dans le champ mots-clés. Ex: https://hal.inrae.fr/hal-03147904v1 https://hal.inrae.fr/hal-03087709v1 https://hal.inrae.fr/hal-03189143v1

Concernant le nombre de descripteurs utilisés, on observe des pratiques très diverses puisqu'il peut varier de 3 à 20.