Révéler et valoriser vos données avec le thésaurus INRAE

Sophie Aubin, Emilie Bernard, Sonia Bravo, Colette Cadiou, Agnès Girard, Magalie Weber

Révéler et valoriser vos données avec le thésaurus INRAE

Sophie AUBIN[1]
Émilie BERNARD[2]
Sonia BRAVO[1]
Colette CADIOU[3]
Agnès GIRARD[4]
Magalie WEBER[5]

Correspondance
sophie.aubin@inrae.fr

Résumé.

Dans un contexte de surabondance de l'information, disposer d'outils d'aide à l'indexation et à la recherche documentaire s'avère de plus en plus indispensable. La construction d'un thésaurus couvrant les thématiques d'un institut de recherche comme INRAE (Institut national de recherche pour l'agriculture, l'alimentation et l'environnement) a pour objectif de décrire finement les productions institutionnelles. Ce type de vocabulaire permet de mieux valoriser ces productions en facilitant leur découverte et leur réutilisabilité. Le thésaurus INRAE contient plus de 16 000 concepts en français et en anglais, des synonymes et des définitions et sert actuellement à indexer des publications, codes et données scientifiques avec des mots-clés. Après avoir présenté le thésaurus INRAE, en détaillant sa structure, son contenu et les divers moyens d'y accéder, nous illustrons son utilité à travers cinq témoignages provenant de collectifs d'utilisateurs scientifiques ou d'appui à la recherche.

Mots-clés

Thésaurus, vocabulaire contrôlé, système d'information, indexation, annotation, terminologie, sémantique, interopérabilité

1 INRAE, DipSO, F-49070 Beaucouzé, France.
2 INRAE, PEGASE, Institut Agro, F-35590 Saint-Gilles, France.
3 INRAE, DipSO, F-63172 Aubière, France.
4 INRAE, LPGP, F-35000 Rennes, France.
5 INRAE, BIA, F-44000 Nantes , France.

Reveal and boost your data with the INRAE thesaurus

Sophie AUBIN[1]
Émilie BERNARD[2]
Sonia BRAVO[1]
Colette CADIOU[3]
Agnès GIRARD[4]
Magalie WEBER[5]

Correspondence
sophie.aubin@inrae.fr

Abstract.

In a context of information overload, it is becoming increasingly essential to have customized indexing and document search tools. The construction of a thesaurus covering the themes of a research institute like INRAE (National Research Institute for Agriculture, Food, and the Environment) aims to provide a detailed description of institutional productions. This type of vocabulary enhances the value of these products by making them more findable and reusable. The INRAE thesaurus contains over 16,000 concepts in French and English, including synonyms and definitions, and is currently used to index publications, codes, and scientific data with keywords. After presenting the INRAE thesaurus, detailing its structure, contents, and various ways of accessing it, we demonstrate its usefulness through five examples of feedback from scientific or research support user groups.

Keywords

Thesaurus, controlled vocabulary, information system, indexing, annotation, terminology, semantics, interoperability

1 INRAE, DipSO, F-49070 Beaucouzé, France.
2 INRAE, PEGASE, Institut Agro, F-35590 Saint-Gilles, France.
3 INRAE, DipSO, F-63172 Aubière, France.
4 INRAE, LPGP, F-35000 Rennes, France.
5 INRAE, BIA, F-44000 Nantes , France.

Introduction

Dans le contexte de la Science Ouverte, INRAE a développé son propre thésaurus afin d'améliorer la visibilité de ses productions scientifiques ainsi que l'interopérabilité de ses systèmes d'information entre eux et avec l'extérieur. Mettre en place l'interopérabilité sémantique[6], c'est permettre la compréhension des données, de façon non ambiguë, en se basant sur des vocabulaires contrôlés et partagés au sein d'un ou de plusieurs domaines scientifiques. Le thésaurus INRAE permet à l'Institut de disposer d'un référentiel terminologique adapté à ses besoins, standardisé et évolutif. Le thésaurus INRAE est en effet le premier vocabulaire, ouvert et partagé sur le web, couvrant tous les domaines de recherche INRAE. Riche de 16 000 concepts bilingues, en français et en anglais, il permet de décrire tous types de productions scientifiques : documents, images, jeux de données, pages web, descriptions d'activités... Cela permet donc d'alimenter un moteur de recherche multilingue, des outils d'analyse de l'information ou d'aide à la traduction.

Cet article présente la ressource, sa structure, ses usages ainsi que l'intérêt de son utilisation dans un projet scientifique, pour un site web ou un système d'information. Cinq témoignages illustreront ces applications.

Présentation du thésaurus

Un thésaurus est un ensemble organisé de termes qui expriment les concepts utiles à la description de contenus propres à un domaine de connaissance. Il permet de gérer le multilinguisme et la synonymie. C'est un outil documentaire, généralement utilisé pour indexer des contenus à l'aide de mots-clés. Partagé par plusieurs systèmes d'informations, il permet d'interconnecter les ressources qui y sont hébergées.

La création du thésaurus INRAE s'inscrit dans la démarche Science Ouverte[7] de l'Institut et la mise en œuvre des principes FAIR[8]. Cela permet d'avoir un référentiel standardisé et évolutif, couvrant l'ensemble des thématiques scientifiques de l'Institut. Celui-ci est validé par la communauté scientifique INRAE et destiné à répondre en première intention aux besoins de l'Institut.

Le thésaurus INRAE, mis en service en mars 2021, est administré par la Direction pour la Science Ouverte (DipSO) avec le concours du réseau de professionnels de l'information scientifique et technique (IST). Le comité éditorial compte vingt contributrices et contributeurs réguliers qui peuvent faire appel à des scientifiques pour la création et la mise à jour de concepts notamment via des relectures ponctuelles de certaines parties du thésaurus.

Comment est structuré le thésaurus ?

Le thésaurus INRAE est le résultat de la fusion de l'ancien thésaurus IRSTEA[9] (8 000 termes) et du référentiel mots-clés INRA (20 000 termes) réalisée en 2020. Sa conception est conforme aux normes internationales relatives aux thésaurus et à leur interopérabilité (Normes ISO 25964 T.1 2011 et T. 2 2014). Il est représenté à l'aide du standard SKOS (Simple Knowledge Organisation System, recommandation W3C 2009) qui facilite sa réutilisation (Yon et Aubin, 2022). Sa structure a été conceptualisée et complétée à partir d'un travail d'analyse comparative avec d'autres vocabulaires (AGROVOC, GEMET, MeSH, thésaurus de la biodiversité...).

Le thésaurus est organisé de manière thématique, de façon à représenter les objets de recherche et les méthodes employées à INRAE. Il est composé d'environ 16 000 concepts, répartis dans 12 domaines et 63 microthésaurus (MT) (Figure 1).

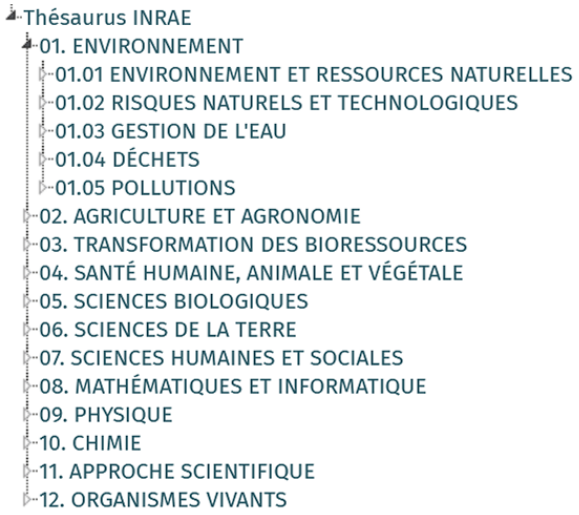

Figure 1. hiérarchie du thésaurus et ses 12 domaines, et les 5 microthésaurus du domaine Environnement (lien vers l'image png : https://science-ouverte.inrae.fr/fr/le-numerique-pour-la-science-et-les-donnees-scientifiques/produire-des-donnees-fair).

Les microthésaurus constituent des ensembles thématiques de concepts organisés entre eux par des relations hiérarchiques.

En figure 2, le concept « algue » est le concept parent de 4 concepts fils.

6 https://openscience.pasteur.fr/2022/06/03/comment-rendre-ses-donnees-interoperables/.

7 http://science-ouverte.inrae.fr/fr/la-science-ouverte/la-politique-de-science-ouverte-dinrae.

8 http://science-ouverte.inrae.fr/fr/les-donnees-et-le-numerique-scientifiques/produire-des-donnees-fair.

9 INRAE est issu de la fusion de IRSTEA et l'INRA en 2020.

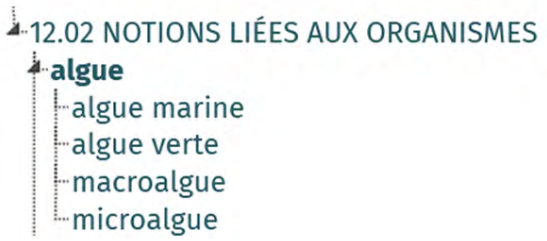

Figure 2. exemple de hiérarchie pour le concept « algue »

Les concepts peuvent aussi être regroupés au sein de collections. Cela permet une autre visualisation des concepts sous forme de listes pour répondre à des besoins spécifiques, par exemple sur une thématique : Bois et forêt, ou une propriété : concepts portant une définition. Ces collections peuvent servir de vocabulaire contrôlé d'un système d'information particulier.

Chaque concept, par exemple ici « agroécologie » (Figure 3), contient un terme préférentiel en français, parfois des synonymes, une définition et une ou plusieurs traductions en anglais. Les définitions proviennent de sources validées (Aubin et al. 2020 ; Aubin, 2021), et de préférence produites par les scientifiques de l'Institut afin d'assurer une bonne adéquation avec les besoins en interne et aussi de valoriser les connaissances produites à INRAE.

Le comité en charge de la maintenance du thesaurus a élaboré un guide à usage interne qui fixe les règles éditoriales relatives à la structure et aux contenus pour assurer le développement homogène du thésaurus. Ce guide contient une description thématique de chaque microthésaurus. Il explicite les décisions prises pour le classement des concepts et les grandes règles à appliquer (variantes orthographiques, acronymes, synonymes...).

Relations avec d'autres vocabulaires

Le thésaurus est aligné sur différents vocabulaires produits en interne (French Crop Usage) ou par des tiers (AGROVOC, GEMET, MeSH). En créant une équivalence entre deux concepts représentant la même chose dans deux ressources différentes, les alignements permettent de passer d'un vocabulaire à l'autre. Dans la figure 3, le concept « agroécologie » est aussi présent dans le thésaurus Agrovoc (GR

Figure 3. Affichage d'un concept dans le thésaurus (lien vers l'image png : https://nextcloud.inrae.fr/s/zzPxjDdHmNdrfwz)

alignés à Agrovoc). Ces alignements contribuent à l'interopérabilité sémantique des systèmes d'information permettant ainsi un échange de données sans ambiguïté.

À quoi et à qui peut servir le thésaurus INRAE ?

Le thésaurus INRAE permet d'indexer et d'annoter tous types de productions scientifiques : documents, images, jeux de données, codes, pages web… Il peut potentiellement servir à alimenter un moteur de recherche multilingue, des outils d'analyse de l'information ou d'aide à la traduction. Le thésaurus peut également être utilisé comme vocabulaire pivot. Un vocabulaire pivot, par le jeu des synonymes et des traductions, permet de relier des données ou des informations équivalentes mais exprimées différemment.

La figure 4 illustre des usages du thésaurus pour divers types d'utilisateurs internes et externes à INRAE tels que des scientifiques, documentalistes, gestionnaires de sites web, concepteurs de système d'information, gestionnaires ou curateurs de données.

Comment consulter le thésaurus INRAE ?

Le thésaurus INRAE est consultable depuis le portail : https://consultation.vocabulaires-ouverts.inrae.fr/thesaurus-inrae/fr/ et, depuis février 2024, sur AgroPortal, le portail des ressources sémantiques en agriculture et domaines associés : https://agroportal.lirmm.fr/ontologies/INRAETHES.

Glossaire :

Terme préférentiel	Libellé choisi pour caractériser le concept
Concept(s) associé (s)	Concept voisin, relation d'association « voir aussi »
Synonyme (s)	Libellé non utilisé au profit du terme préférentiel (ex : synonyme, acronyme...)
Appartient aux groupes	Les groupes sont des regroupements de concepts
Microthésaurus	Partie thématique du thésaurus à laquelle appartient le concept
Traduction(s)	Traduction du terme et de ses synonymes
URI	Uniform Resource Identifier : lien vers le concept

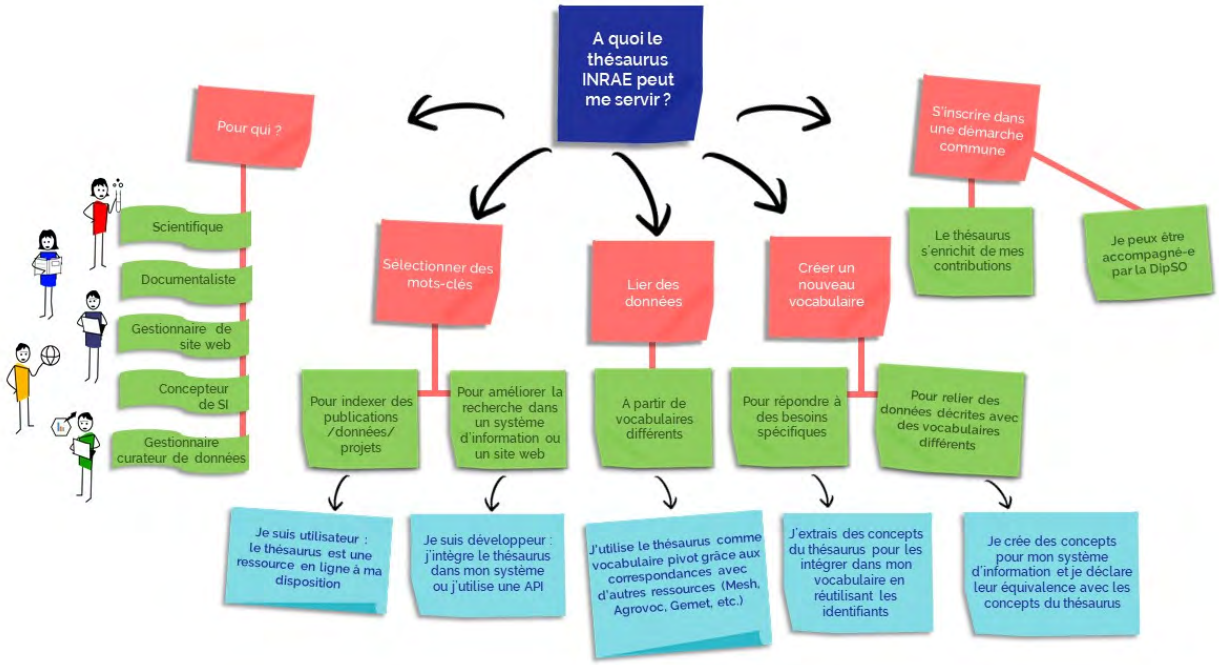

Figure 4. Usages potentiels du thésaurus

Deux API[10] permettent à des applications d'exploiter le thésaurus : l'API PADRE, destinée à des applications internes INRAE puisqu'elle nécessite une authentification, et l'API Skosmos qui est accessible librement.

Témoignages d'utilisateurs INRAE

Au sein d'INRAE, plusieurs projets et applications avec des périmètres et thématiques variés se sont emparés de cette ressource terminologique. Les cinq témoignages d'utilisateurs présentés ci-dessous illustrent tout l'intérêt du thésaurus pour la valorisation des travaux de recherches de l'Institut.

Une indexation fine et pertinente des publications scientifiques

Témoignage de Clotilde Nicol, Isabelle Massart, Vincent Rappeneau (DipSO).

« Hal INRAE (https://hal.inrae.fr) est l'archive ouverte destinée au dépôt et à la consultation des travaux scientifiques de l'Institut. L'intégration du thésaurus INRAE dans l'offre de vocabulaires existante de HAL (MeSH, JEL, les classifications ACM…) est venue répondre aux besoins d'indexation de documents sur nos thématiques de recherches. Le déposant peut interroger le thésaurus à partir d'un champ du formulaire qui fonctionne par auto-complétion. Le document est non seulement indexé avec le concept sélectionné mais aussi avec l'ensemble des termes qui le définissent (synonymes, traductions). Ainsi le document est plus facilement trouvable dans HAL : que la recherche soit lancée sur l'un ou l'autre des termes du concept qui l'indexe, le moteur le retourne dans sa réponse. Cette propriété fait du thésaurus INRAE un puissant outil d'indexation. Pour un usage optimal, une politique d'indexation a été arrêtée. Elle est consultable ici : https://hal.inrae.fr/hal-033469n40. »

Des jeux de données plus faciles à trouver

Témoignage de Cathy Tang (DipSO).

« Data INRAE[11], un des espaces institutionnels de l'entrepôt de données Recherche Data Gouv, permet le partage et la valorisation des données de recherche de l'Institut.
Pour favoriser la compréhension et la visibilité de ces données, il est important de bien les décrire en renseignant notamment des mots-clés. La récupération de mots-clés issus d'un vocabulaire ouvert reconnu par la communauté, comme l'est le thésaurus INRAE, participe aux principes FAIR en favorisant notamment l'interopérabilité avec d'autres systèmes.

10 API : Application programming interface ou « interface de programmation d'application ».
11 https://entrepot.recherche.data.gouv.fr/dataverse/inrae.

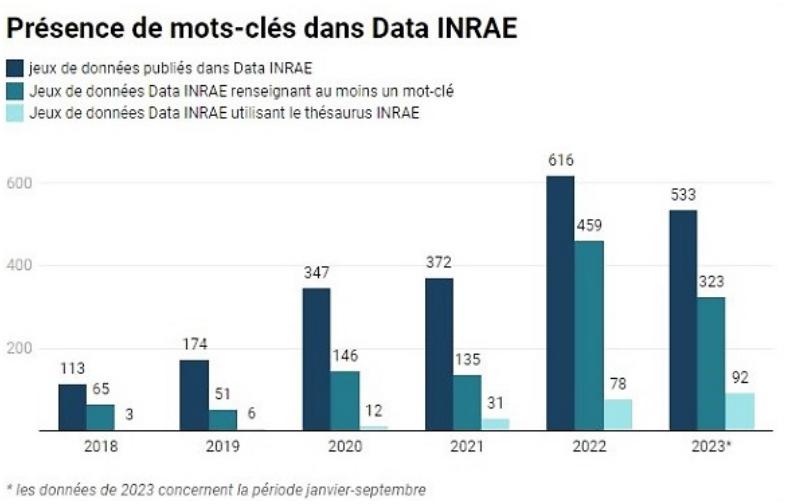

Figure 5. présence des mots-clés dans Data INRAE (lien vers image png : https://nextcloud.inrae.fr/s/EPKEZaDkXLWzXHP)

SICPA : Décrire les variables d'un système d'information

Témoignage de Bernadette Urban (UMR Pegase) et François Laperruque (UMR Genphyse)

« Le Cati SICPA (Centre automatisé pour le traitement de l'information sur les Systèmes d'informations et calcul pour le phénotypage animal) propose des outils de collecte et de stockage d'informations issues des unités et installations expérimentales animales des départements GA (Génétique animale) et Phase (Physiologie animale et systèmes d'élevage). L'intégration du Cati à l'infrastructure de recherche LiPh4SAS et son inscription dans la démarche d'ouverture des données ont nécessité la constitution d'un catalogue des données, produites et gérées en interne.

Nous avons établi avec les chefs de projets informatiques de tous nos systèmes d'information, un premier niveau de description pour aboutir à une version 1 de ce catalogue qui comporte 718 types de données différents (des phénotypes aux codes sources).

Le thésaurus INRAE est un référentiel très complet, qui nous permet de pratiquement décrire toutes nos données grâce à un terme ou une combinaison de deux ou trois d'entre eux. Il est géré par un collectif d'une grande réactivité, tant pour les conseils de recherche que pour l'évolution du référentiel par l'ajout de nouveaux éléments.

Nous visons une description complète des variables de nos systèmes d'information au travers de ce référentiel avec un grain de description qui devrait progressivement s'affiner. ».

Pour en savoir plus : voir le plan de gestion de données du Cati SICPA : https://doi.org/10.57745/S19P9X.

MAGGOT : une aide à la bonne gestion des données avec un vocabulaire spécifique

Témoignage de Daniel Jacob (INRAE UMR BFP) et François Ehrenmann (INRAE UMR BioGECO).

MAGGOT est un outil de gestion des métadonnées qui répond aux problèmes d'organisation, de documentation, de stockage et de partage des données dans une unité ou une infrastructure de recherche, et s'intègre parfaitement dans un plan structurel de gestion des données. Il est développé dans les unités BFP (Biologie du Fruit et Pathologie) et BioGECO (Biodiversité Gènes et Communautés) avec le concours des CATI PROSODIe et GEDEOP. En savoir plus sur Maggot : https://inrae.github.io/pgd-mmdt/.

« Annoter ses données requiert de pouvoir disposer d'un vocabulaire contrôlé adapté à son domaine d'application. Cela peut impliquer l'utilisation de vocabulaire métier, donc très spécifique, qui reflète une réalité complexe et/ou très précise. C'est pourquoi, nous avons fait le choix de l'utilisation de thésaurus, tel le thésaurus INRAE, dans notre outil Maggot servant à annoter des jeux de données en créant un fichier de métadonnées à joindre dans l'espace de stockage. »

Info&Sols : Des mots-clés pertinents pour la diffusion de données publiques

Témoignage de Christine Le Bas, unité Info&Sols.

« L'unité INRAE Info&Sols est en charge, au nom du groupement d'intérêt scientifique sur les sols (GIS Sol), de la diffusion des données publiques du système d'information sur les sols de France. Cette diffusion s'effectue dans différents portails selon la nature des données mais tous nécessitent l'usage de mots-clés. Nous utilisons le thésaurus INRAE car il est plus adapté à notre domaine. Il nous permet ainsi de mieux décrire nos jeux de données et d'homogénéiser une grande partie des mots-clés que nous utilisons ce qui améliore la recherche et le catalogage des jeux de données. Le caractère bilingue du thésaurus nous permet également d'améliorer l'internationalisation de nos jeux de données. Nous collaborons également avec l'équipe chargée du thésaurus pour en améliorer le contenu et avoir une plus grande adaptation du thésaurus à nos besoins. »

Pour en savoir plus : https://info-et-sols.val-de-loire.hub.inrae.fr/

Conclusions et perspectives

Comme le montrent les témoignages précédents, le thésaurus INRAE est un outil de référence essentiel et simple d'utilisation pour indexer diverses productions scientifiques à l'aide de mots-clés, ce qui améliore la performance de la recherche bibliographique. Il peut aussi servir de base de référence pour l'aide à la traduction français-anglais, pour la création de vocabulaires spécifiques, ou encore la description fine de variables ou de réalités complexes. Des développements se poursuivent pour intégrer le thésaurus dans diverses plateformes, à l'instar de Data INRAE où il est désormais disponible.

Pour répondre au mieux à ces besoins variés, la DipSO maintient cette ressource collective avec l'appui du réseau de professionnels de l'IST et la validation de scientifiques. Le thésaurus s'enrichit ainsi au fil de l'eau des remarques, commentaires et demandes d'évolution des différents utilisateurs[12]. Les nouveaux contributeurs sont les bienvenus de manière ponctuelle ou sur la durée.

Pour une meilleure appropriation du thésaurus, le comité éditorial devra identifier et analyser les freins et opportunités concernant son utilisation. Un plan de communication et de formation est mis en place pour inciter la communauté INRAE à développer et diversifier les usages du thésaurus. En outre, l'affichage du thésaurus sur le portail international AgroPortal permet d'accroître sa visibilité en dehors de l'Institut et d'attirer ainsi de nouveaux utilisateurs ou de faire remonter de nouveaux besoins internes ou externes.

Si vous souhaitez partager votre expérience d'utilisation du thésaurus avec la communauté, ou faire part de besoins particuliers, le comité éditorial[13] est à votre disposition. ■

Contact du comité éditorial du thésaurus INRAE :
thesaurusINRAE@inrae.fr

Remerciements

Merci à nos relectrices : Valérie Pagneux, Marie-Pierre Maleyran-Raymond, Pascale Hénaut.

12 Entre mi-2021 et mi-2024, soit en 3 ans, le thésaurus s'est enrichi de 643 concepts, 900 définitions et 4500 traductions en anglais.

13 Composition du comité éditorial et liste des contributeurs : consulter le site Vocabulaires ouverts INRAE : https://vocabulaires-ouverts.inrae.fr/a-propos-du-thesaurus-inrae/.

Références

Aubin S., Cadiou C., Weber M. (2020) Intégrer des définitions textuelles à mon thésaurus ou à mon ontologie. Comment rester dans la légalité ? INRAE. 4 p. doi:10.15454/sbey-xa19.

Aubin, S. (2021). Rédiger une définition, quelques clés. INRAE. 8 p. doi:10.15454/93xg-p646.

International Organization for Standardization (2011) Information et documentation - Thésaurus et interopérabilité avec d'autres vocabulaires - Partie 1 : thésaurus pour la recherche documentaire (ISO Standard 25964-1:2011).

International Organization for Standardization (2013) Information et documentation - Thésaurus et interopérabilité avec d'autres vocabulaires - Partie 2 : interopérabilité avec d'autres vocabulaires (ISO Standard 25964-2:2013).

SKOS : Système Simple d'Organisation de Connaissances Référence Recommandation W3C du 18 août 2009. Disponible sur : http://www.sparna.fr/skos/SKOS-traduction-francais.html, traduction de : http://www.w3.org/TR/2009/REC-skos-reference-20090818/ ; dernière mise à jour : avril 2014.

Yon J., Aubin S. (2022) SKOS : un standard pour une ressource sémantique simple et FAIR. Disponible sur : https://vocabulaires-ouverts.inrae.fr/skos-un-standard-pour-une-ressource-semantique-simple-et-fair/.