Amélioration du Thesaurus VOCINRA en utilisant les standards du web sémantique

Fama Diop

AMÉLIORATION DU THESAURUS VOCINRA EN UTILISANT DES OUTILS ET STANDARDS DU WEB SÉMANTIQUE

FAMA DIOP

Université Paris Sorbonne Nouvelle

Mémoire de master 2

Établissements : Sorbonne Nouvelle, Paris Nanterre la Défense, Inalco

Mention : Traitement automatique des langues

Spécialité : Recherche & Développement

Année universitaire 2015-2016

Mémoire dirigé par Jean-Luc Minel

Stage encadré par Pascal Aventurier et Sophie Aubin

Remerciements

Je voudrais remercier toutes les personnes de près ou de loin qui ont participé à la réalisation de ce travail.

Je tiens à remercier M. Jean-Luc Minel pour sa confiance , ses contributions, ses conseils et surtout aussi pour ses enseignements qui m'ont été vraiment utiles tout au long de ce travail .

Merci aussi Mme Delphine Battistelli d'avoir accepté de faire parti du jury.

J'adresse un remerciement particulier à M. Pascal Aventurier et Mme Sophie Aubin de m'avoir donné l'opportunité de travailler avec eux. Je ne pourrais jamais vous remercier assez pour les conseils, les remarques, le soutien et les nouvelles connaissances acquises. Sachez que vous êtes plus que des tuteurs pour moi. Merci aussi à M. Philippe Clastre qui a toujours été là quand VocBench ne marchait pas.

Je remercie aussi les enseignants du Master TAL qui nous ont accompagnés durant ces deux années d'étude particulièrement à M. Serge Fleury pour sa disponibilité et sa réactivité.

Je ne remercierai jamais assez mes parents pour tout ce qu'ils ont fait et continuent de faire pour moi malgré la distance, ils sont toujours présents à travers leur soutien et conseil. Je remercie aussi mes frères et sœurs pour tout ce que nous avons partagé et continuons de partager. L'amour qu'on se porte est plus fort que la distance qui nous sépare, chère famille.

Je ne pourrai terminer sans remercier un oncle M. El Hadji Mbaye qui m'a énormément soutenu durant tout mon cursus scolaire.

Sommaire

1. Introduction

Ce mémoire portant sur le thésaurus VocINRA (vocabulaire de l'INRA) a été rédigé dans le cadre de mon stage au pôle de Gestion des Connaissances (GeCo) de l'INRA. L'INRA est l'Institut National de la Recherche Agronomique pour la France, leader de son domaine en Europe et second en nombre de publications scientifiques à l'échelle mondiale. Le pôle GeCo est un des pôles composant le Département de l'Information Scientifique et Technique (DIST) qui permet à un organisme de regrouper, d'analyser et de valoriser l'ensemble des connaissances produites par les chercheurs de l'INRA ou qui leur seront utiles dans leur recherche. Le pôle GeCo se fixe pour objectif d'accompagner les agents ou les équipes de l'INRA pour construire ou publier leur vocabulaire représenté dans des thésaurus ou ontologies. Pour l'élaboration de tels vocabulaires, le pôle intervient notamment dans le choix des outils ou standards adoptés comme OWL pour les ontologies et Skos ou Skos-XL pour les thésaurus. Ces standards sont des recommandations dans le cadre du web sémantique pour pouvoir faciliter le partage et l'interopérabilité de telles données sur le web.

L'un des objectifs du projet VocINRA est de pouvoir améliorer le vocabulaire de utilisé pour l'indexation des publications de l'INRA en commençant par le format de base qui est du XML pour le transformer en du Skos. En plus, comme le vocabulaire est managé par un outil qui n'est pas destiné à du Skos. Il faudra à terme le remplacer par un autre qui accepte ce format. Pour le choix de l'outil, comme notre vocabulaire est destiné à l'agronomie et aux thèmes en rapport avec, nous avons décidé de tester VocBench qui est utilisé pour la gestion des thésaurus du même domaine à savoir GACS et AGROVOC.

Dans un premier temps, nous allons faire une présentation plus approfondie sur les objectifs à atteindre dans ce projet et parler de quelques approches qui ont été développées à propos de la gestion des connaissances et aussi des recommandations du web sémantique, domaines auxquels ce travail est lié.

Nous présenterons ensuite les définitions données à un thésaurus ainsi que les normes établies pour son élaboration.

Les informations constituant un thésaurus doivent être évaluées pour vérifier si la qualité répond aux exigences fixées ou pas. Donc, nous verrons cela une fois que nous aurons présenté les méthodes qui nous ont permis de transformer nos données en Skos.

A travers ce travail, nous montrerons aussi comment le traitement automatique des langues nous a été utile pour réaliser nos tâches de filtrage (cinquième partie) et d'extraction des synonymes (sixième partie).

Nous allons aborder aussi en dernier lieu un élément important dans ce projet qui est l'alignement à faire entre VocINRA et les thésaurus GACS et AGROVOC. L'alignement des concepts entre thesauri permet d'identifier les concepts communs entre eux afin de récupérer par exemple des informations comme les labels de concepts dans d'autres langues

1.1 Objectifs

Le thésaurus VocINRA sert à l'indexation de l'Archive Ouverte ProdINRA (http://www.ProdINRA.fr), spécialisée en agronomie et domaines associés tels que l'environnement, les sciences sociales, la génomique, les biosciences, etc. Il est multilingue, principalement en français et anglais.

VocINRA a été constitué à partir de différentes sources et est en constante évolution puisque les documentalistes de l'Inra proposent de nouveaux mots clés selon un workflow de validation/enrichissement. Des travaux d'alignement de VocINRA avec des thésaurus de référence comme AGROVOC/GACS ont démarré et seront poursuivis, en particulier dans le cadre d'une collaboration avec l'Embrapa Brésil. La publication d'une version en Linked Open Data est également prévue.

VocINRA est actuellement maintenu et exploité à l'aide d'un outil développé spécifiquement en interne et ne prenant en charge que des fichiers au format XML. Ce format n'est pas recommandable pour la gestion des thésaurus interopérables pour le web sémantique qui privilégie le SKOS. Un thésaurus en Skos distingue les termes et les concepts or dans notre thésaurus tout est termes. Donc, les termes partageant le même réseau sémantique sont à regrouper autour d'un même.

L'outil de gestion est à remplacer par un autre qui permettra de faire évoluer le vocabulaire de l'éditer en gestion de projet pour que d'autres personnes collaborent à son management. VocBench (http://aims.fao.org/fr/vest-registry/tools/VocBench-2) offre la possibilité d'éditer son thésaurus en gestion de projet. Nous allons ainsi le tester pour savoir s'il sera possible pour nous de remplacer l'IMC par VocBench. Un des objectifs de tester cet outil est de savoir si nous allons pouvoir réorganiser l'arborescence de notre thésaurus, c'est-à-dire essayer de regrouper autour d'un même concept les termes synonymes ou quasi-synonymes.

Comme nous envisageons aussi d'enrichir les concepts avec des liens vers des concepts dans d'autres ressources comme AGROVOC ou GACS, nous verrons si VocBench sera efficace pour réaliser une telle tâche. Si non, essayer de trouver les méthodes qui ont permis aux deux vocabulaires cités de s'aligner entre eux.

En plus, la qualité des données est aussi à étudier comme notre thésaurus s'est constitué majoritairement à partir des indexations réalisées dans ProdINRA. Il faudra s'intéresser à la casse qui n'est pas souvent respectée, vérifier si les fichiers transformés ne contiennent pas des anomalies, si nos données répondent aussi aux règles établies par la Norme 25964 pour pouvoir envisager des corrections.
A la suite de nos réalisations, des recommandations seront à prévoir pour le suivi du projet.

1.2 Gestion des Connaissances

La gestion des connaissances ou Knowledge Management est une méthode qui regroupe les techniques managériales en vue d'identifier l'information, de l'organiser, de l'analyser et de la diffuser au sein de l'entreprise. Ces connaissances sont souvent des informations qui sont produites par l'organisation en question ou par d'autres entreprises. D'après Jean-Louis Ermine, [Emine 2008], qui s'est lui aussi basé sur les théories de Nonaka et Takeuci, il existe deux types de connaissances : les connaissances tacites qui sont personnelles c'est-à-dire elles proviennent de l'individu et les connaissances explicites qui sont interprétables par l'entreprise. Pour gérer ces connaissances, une entreprise doit s'appuyer sur des différents outils et technologies comme les messageries, les agendas, le groupware, le workflow, la gestion documentaire, les moteurs de recherche, les outils de veilles, les cartographies, les portails etc. Comme l'ère des papiers est un peu révolu et que nous sommes à l'ère du numérique, ces connaissances sont souvent managées à l'aide de l'informatique. Si nous prenons le cas de la gestion électronique des documents, les documents sont indexés et ces indexations se font à l'aide de thésaurus qui ont été constitués au préalable. Nous nous rendons compte que ceci est un moyen de classifier les savoirs qui ont déjà été collectés. Par ailleurs, pour le partage et l'interopérabilité de ces informations il faut des techniques qui permettent d'y arriver. Donc, il faut que les connaissances collectées soient dans un format destiné à simplifier ce partage et pour cela il faut que les données soient modélisées aussi. Pour modéliser des données et les partager sur le web, des recommandations ont été développées par le W3C pour le web sémantique ou web des données.

1.3 Web Sémantique, Web de données

1.3.1 Web Sémantique

Le Web actuel est composé d'un ensemble de données dont l'information n'est pas assez structurée n'est pas structurée du tout malgré quelques tentatives de balises méta et de micro

format et est incompréhensible par les machines car il est purement syntaxique et pauvre en sémantique ; les données et les métadonnées ne sont pas interprétables par les machines. Donc elles sont incapables de comprendre le sens des contenus d'une page Web pour déduire (inférer) de nouvelles connaissances, proposer des contenus similaires au lecteur, agréger des contenus, etc. Faire des recherches est possible grâce aux moteurs de recherche qui indexent le texte intégral grâce à des algorithmes sophistiqués. Malheureusement, les réponses fournies ne sont pas souvent celles attendues par l'utilisateur. En général, ces réponses ne sont pas bien organisées ce qui fait que c'est l'utilisateur qui doit chercher à son tour parmi les informations proposées celles qui correspondraient à sa demande. Donc, le tri de l'information n'est pas fait par les machines mais par l'humain.

Pour remédier à ces limitations qui font que le Web est juste assimilé à une base de stockage de documents, une nouvelle structure se verra proposée : le Web sémantique.

Expression inventée par Tim Berners-Lee, Co-inventeur du Web, le Web sémantique est une extension du Web d'aujourd'hui et est un moyen de rendre les machines capables de comprendre le contenu des informations qui composent le Web, c'est dans ce sens qu'il fait cette déclaration lors d'un entretien : « J'ai un double rêve pour le Web. D'une part, je le vois devenir un moyen très puissant de coopération entre les êtres humains. Et dans un second temps, j'aimerais que ce soit les ordinateurs qui coopèrent. [...] Quand mon rêve sera réalisé, le Web sera un univers où la fantaisie de l'être humain et la logique de la machine pourront coexister pour former un mélange idéal et puissant. »

Mettre en place une nouvelle architecture, va "amener le Web à exploiter son vrai potentiel" [Studer 2003]. Les informations seront sémantiquement structurées et stockées dans des bases de connaissances définies sous forme des concepts et de relations [Gandon 2012]. Ainsi, les machines seront en mesure de fournir des réponses pertinentes aux recherches des utilisateurs et le Web deviendra un guide intelligent en traitant des informations qu'il comprendra comme les humains.

Ces technologies sont souvent représentées à l'aide d'une pyramide comme l'a proposé pour la première fois Berners-Lee en 1998.

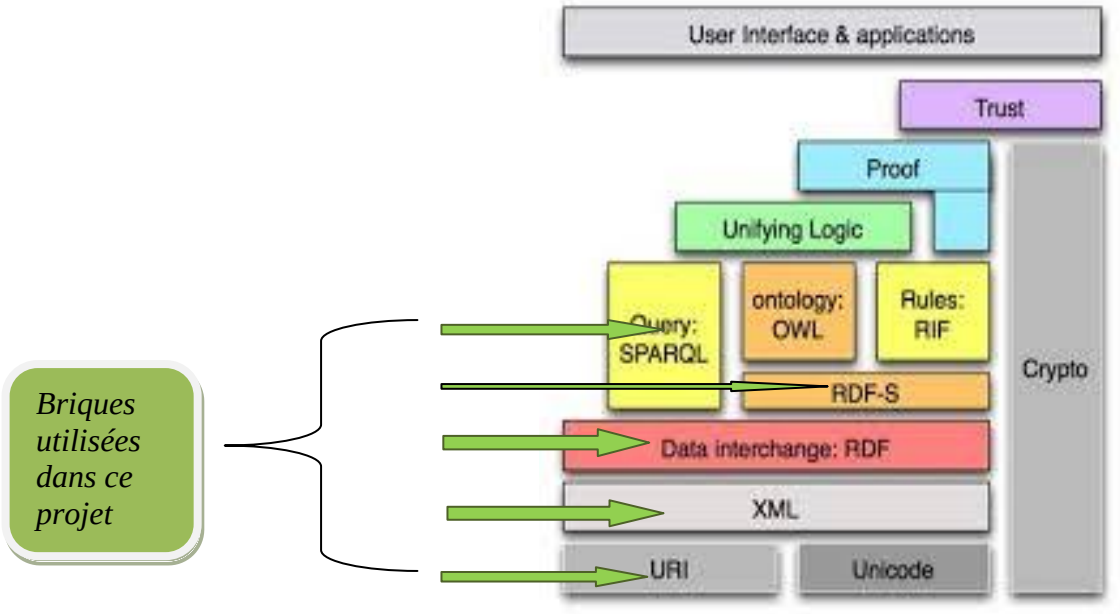

Figure 1 pyramide du web sémantique et les parties utilisées dans notre projet

Ces briques qui forment la pyramide des recommandations du W3C ne sont pas intimement liées les unes aux autres, car elles sont indépendantes. Chaque brique peut être utilisée sans pour autant faire appel à une autre brique.

L'architecture du Web d'aujourd'hui repose sur trois notions fondamentales . L'URL (Uniform Resource Locator) qui est une chaîne de caractères et un sous-ensemble des URI permet d'allouer une adresse aux ressources sur le Web pour pouvoir les identifier ; le protocole HTTP (HyperText Transfer Protocol) a pour but de permettre à un utilisateur d'avoir accès au serveur dans lequel les données se trouvent ; le langage HTML (HyperText Markup Language) est un langage à balises permettant de produire des documents hypertextuels appelés page Web qui pourront être visualiser par les utilisateurs.

Si le Web d'aujourd'hui repose sur les trois notions citées, le Web sémantique comme constaté sur la pyramide a pour première brique l'URI qui est un moyen d'identifier n'importe quelles données dont on se réfère et l'Unicode renvoie au standard informatique utilisé pour encoder les données. Le protocole HTTP se chargera toujours du transfert des données [Gandon 2012].

Un nouveau langage nommé RDF (Resource Description Framework) sera proposé à la place du HTML pour pouvoir exploiter les données, les informations et les métadonnées. Les métadonnées sont des données qui décrivent d'autres données. Elles sont ces marqueurs introduits dans les fichiers et également dans les langages de programmation. Dans le cadre du web sémantique, ce sont dans les pages web que sont contenues les métadonnées en plus du contenu textuel. Donc l'idéal serait que ces métadonnées puissent permettre aux ressources informatiques d'être interopérables et que son rôle ne puisse plus se limiter qu'à la structuration des pages web, ce qui donnera aux agents logiciels l'accès à ces métadonnées

afin qu'elles soient exploitables par eux. L'utilisation des métadonnées apporte également une meilleure structuration dans les classifications comme les thésaurus ou les taxonomies, ce qui est le cas dans notre thésaurus. Par ailleurs, l'utilisation de ces métadonnées dans le cadre du web sémantique a mené à l'élaboration de 15 normes autour du Dublin Core pour faciliter la création des notices descriptives. Dans le cadre de VocINRA ces normes seront représentées dans des fichiers RDF pour structurer les informations portant sur l'auteur d'un mot-clé, sa date de création ou le sujet auquel il appartient par exemple.

Pour faire une description avec le langage RDF, certains éléments sont nécessaires pour y parvenir à savoir : les **ressources**, les **triplets**, et les **propriétés**. Une ressource renvoie à un objet qui décrit une connaissance. Elle est un ensemble d'information classé dans un bloc avec un identifiant nommé URI ou un littéral, une propriété est un caractère spécifique, une relation décrivant une ressource, un triplet est l'association d'un sujet (ressource à décrire), d'un prédicat (propriété applicable à une ressource) et d'un objet (valeur de la propriété).

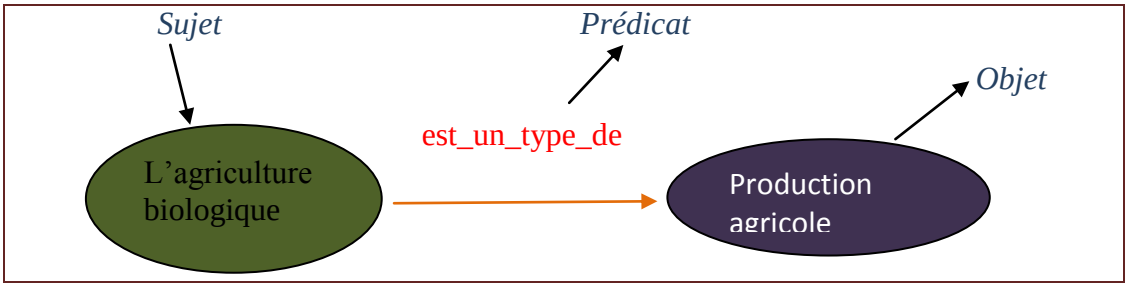

Figure 2 Exemple d'un triplet RDF

Le RDFs (RDF schéma) est un langage extensible du RDF et il permet de représenter des connaissances tout en offrant une meilleure structuration des données. Il est utilisé dans la création des ontologies dites légères pour pouvoir définir les classes et les propriétés tout en offrant la possibilité de les représenter de façon hiérarchisée.

Les connaissances sont représentées dans des ontologies avec différents langages en fonction de leur simplicité ou de leur complexité. Même si les ontologies sont au centre du Web sémantique il n'en demeure pas moins que d'autres langages ont été développés pour traiter du contenu dans le Web comme le SKOS qui est une autre recommandation du W3C parue en

2009[W3C]. SKOS (Simple Knowledge Organization System) toujours fondé sur le RDF est le langage qui permet de représenter les thésaurus, les vocabulaires contrôlés afin d'offrir une meilleure structuration des vocabulaires. SKOS présente un modèle de données qui organise les connaissances sous un schéma de concepts qui est « [1]une agrégation de concepts ». Il permet de regrouper les termes d'une même catégorie sous un concept (Skos :Concept) et de définir les relations entre des concepts ou des termes (Skos :narrower, Skos :broader, Skos :hasTopConcept). Les libellés lexicaux bénéficient aussi d'un vocabulaire (Skos : prefLabel, Skos : altLabel, Skos : hiddenLabel).

Skos sera le standard qui va plus nous intéresser car l'un des objectifs de ce travail est de transformer nos données qui sont au format XML en du SKOS. Les étapes à suivre pour y arriver seront détaillés dans la troisième partie du travail.

SPARQL est une des recommandations du W3C. C'est un langage qui permet d'interroger les données en RDF, de les modifier ou les supprimer sous forme de graphe. Il est un peu proche du SQL langage de requête des bases de données relationnelles.

1.3.1.1 Web de données

Dans cette partie nous allons présenter la notion du web de données qui est un des objectifs à atteindre dans le cadre du projet VocINRA.

Le web de données ne relie pas des documents mais des données en s'appuyant sur les technologies du web sémantique pour permettre aux machines de les interpréter facilement. Le web de données rendra ainsi possible le partage des données et il sera aussi possible pour d'autres de pouvoir exploiter ces données partagées.

Le Web de données ou Linked Data est proposé par le W3C et a pour objectif de relier les données structurées et les publier sur le Web. Actuellement, les données sont en grande majorité stockées dans des bases de données et ne sont quasiment pas structurées ce qui fait qu'elles ne sont pas interopérables. Donc elles ne sont pas liées les unes les autres. L'enjeu serait d'établir un lien entre elles pour les transformer en un réseau sémantique collectif qui facilitera l'échange sur les données par plusieurs utilisateurs .

C'est dans cette optique que Tim Berner-Lee a élaboré quatre grands principes qui vont accompagner le web de données :

1. Se servir des URIS pour pouvoir nommer les ressources

[1] http://www.sparna.fr/Skos/SKOS-traduction-francais.html

2. Utilisation des URI HTTP pour identifier les ressources et avoir accès à ces dernières
3. Les standards RDF, SPARQL, doivent être utilisés quand on veut déréférencer un URI et apporter des informations utiles
4. Créer des liens à partir d'autres URIs pour découvrir d'autres ressources sur le web

Figure 3 Illustration des notions de base du linked data

Url de l'image : https://www.w3.org/DesignIssues/LinkedData.html

Le web des données recouvre aussi le concept du Linked Open Data qui opère dans la gestion des données partagées sur le web pour qu'elles soient liées et ouvertes.

1.4 Thésaurus

1.4.1 Définitions

Le langage documentaire est un langage artificiel qui a pour objectif de faciliter la recherche de l'information ou l'indexation des fonds documentaires qui consiste à attribuer des mots-clés à un document en plus de la description tout en se basant sur ce langage ou le langage naturel. Le lange contrôlé regroupe un ensemble de mots choisis pour caractériser un domaine ou plusieurs contrairement au langage naturel qui propose un nombre infini de mots mais avec plusieurs contraintes. Même si le langage documentaire propose un nombre limité de termes, son efficacité repose sur les normes définies lors de leur élaboration pour contourner par exemple l'utilisation de la majuscule pour les noms communs ou le minuscule pour les noms propres, les relations entre les mots sont établies pou éviter la polysémie également. Pour ranger ses termes, le langage documentaire s'appuie sur des outils comme les thésaurus, les listes d'autorité, les classifications à facettes ou hiérarchiques,

L'outil qui nous intéressera ici sera le thésaurus car c'est notre sujet de recherche. Selon l'AFNOR (Association française de normalisation), le thésaurus est un « Langage documentaire fondé sur une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes de la langue naturelle et les relations entre notions par des signes conventionnels ».

C'est un ensemble de mots simples, composés ou expressions issus du langage naturel et reposant sur une structure et une hiérarchie définies pour aider dans la tâche d'indexation des documents et faciliter l'accès à l'information. Souvent il est confondu avec les dictionnaires qui fournissent des définitions or dans un thésaurus les définitions peuvent être omises.

Le thésaurus est conçu pour un but précis et un domaine particulier comme notre thésaurus, VocINRA, développé pour le champ agronomique.

Dans un thésaurus, les termes sont rangés sous deux catégories : les **descripteurs** et les **non-descripteurs**. Les non-descripteurs sont les termes qui ne sont pas maintenus pour l'indexation mais ils peuvent être employés pour la recherche d'information et sont liés aux descripteurs.

La structuration d'un thésaurus ne se fait pas sans tenir compte de certains facteurs. C'est dans cette optique que des normes ont été définies.

1.4.2 Normes

Dans cette partie nous essayerons de faire une brève présentation des normes qui régissent la structuration et la construction d'un thésaurus. En France, pour élaborer un thésaurus, on fait appel à plusieurs normes que nous essayerons de citer chronologiquement.

- **Normes internationales**

Au départ, deux normes internationales servaient à la création et la structuration d'un thésaurus. Ces normes définies par l'Organisation internationale de normalisation sont l'ISO 2788: 1986 et l'ISO 5964 : 1985. La norme ISO 2788 : 1986 caractérise les étapes à suivre pour développer un thésaurus monolingue en évitant l'ambiguïté des langues et en associant des termes aux concepts tandis que la norme ISO 5964 : 1985 traite des thésaurus multilingues avec les mêmes principes de l'ISO 2788.

Ces deux normes régissent les règles qui permettent, dans les thésaurus monolingues ou multilingues, de répertorier les définitions, les symboles et les abréviations, d'établir les

relations sémantiques qu'entretiennent les unités lexicales. Mais d'après Dextre Clarke et al, le problème d'ambiguïté n'était pas résolu et le schéma de données n'était pas défini.[2] Par ailleurs, l'emploi du pluriel ou du singulier y est traité. Pour la norme ISO, le mieux serait par exemple de mettre tous les éléments du corps humain qui sont en double au pluriel comme « **poumons** » et ceux qui sont uniques au singulier comme «**bouche** ».

- **Normes françaises**

Les normes françaises sont définies par l'AFNOR (Association Française de Normalisation) qui est la représentante de l'ISO et du CEN (Comité européen de normalisation) en France. Les règles d'établissement des thésaurus ont été organisées en deux groupes : NF Z47-100 Décembre 1981 pour les thésaurus monolingues et NF Z47-101 Décembre 1990 pour les thésaurus multilingues. Les normes NF Z47-100 Décembre 1981 et NF Z47-101 Décembre 1990 sont une continuité des normes ISO mais ajoutent quelques changements dans l'élaboration des thésaurus de langue française. Le choix du nombre (pluriel ou singulier) à attribuer aux descripteurs a été souligne. Pour l'AFNOR, il faut mettre tous les descripteurs sauf ceux qui s'écrivent naturellement au pluriel comme *sels minéraux* qui est toujours au pluriel ou *Les Etats-Unis* qu'on ne peut pas mettre au singulier au risque de tronquer le sens.

Comme les langues évoluent et leur usage aussi, ces normes à cause de leur ancienneté vont être modifiées pour pouvoir répondre aux nouvelles exigences tant au niveau international que national.

- **Normes américaines**

Adoptée en Juillet 2005, la norme américaine ANSI/NISO Z39.19 est plus récente que les normes ISO et NF. Elle reprend les directives de ces anciennes normes pour les thésaurus mais apporte beaucoup plus de détail dans le choix des termes. Même si la différenciation entre terme et concept n'est pas posée, la notion de concept est définie sous sept types en intégrant les objets physiques, les matériaux, les activités ou processus, les événements, propriétés, disciplines, unités de mesure et noms propres. Le problème du nombre est toujours posé mais certaines modifications sont apportées. Pour ce qui est des parties du corps, la norme américaine privilégie le singulier même pour les éléments en double. Cette norme apporte beaucoup plus de détails sur les relations aussi. Même si notre thésaurus est conçu à

[2]DEXTRE, Clarcke. «Standard Spotlight: From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling. »

partir des normes françaises (ISO **25964**) nous avons jugé utile de présenter ces normes américaines et montrer les évolutions qu'elles ont apportées dans la conception des thésaurus. Elles pourraient aussi nous être utiles pour le traitement des termes qui sont en anglais comme notre thésaurus est multilingue.

- **Normes anglaises**

La norme britannique BS 8723 tout comme la norme AINSI/NISO est une norme qui apporté beaucoup de nouveautés dans le traitement des thésaurus. Elle gère aussi les ontologies, taxonomies et les systèmes vedettes.[3]

- **Norme ISO 25964**

C'est la nouvelle norme élaborée pour les thésaurus, elle est apparue sous deux volets, le premier paru en 2011 pour la recherche d'information et le deuxième paru en 2013 pour l'interopérabilité des vocabulaires contrôlés. Cette norme remplace les quatre premières normes citées ci-dessus et apporte des nouveautés. Les précédentes normes de l'ISO et de l'afnor traitait les thésaurus de manières séparées, c'est à dire les thésaurus monolingues et multilingues avaient chacun ses règles. La nouvelle norme traite les thésaurus en général qu'ils soient monolingues ou multilingues. Donc, ces règles sont conçues pour ces deux thésaurus en même temps.

Les précédentes normes n'ont pas trop réussi à apporter de distinction entre les concepts et les termes, c'est avec cette nouvelle norme que la distinction sera enfin faite.

Les concepts « peuvent être : des objets et leurs caractéristiques physiques, des matériaux, activités et processus, des événements et faits, des caractéristiques (propriétés) de personnes, objets, matériaux ou actions, des disciplines ou domaines de spécialité (*subject field*), des unités de mesure, types de personnes et organismes ainsi que des entités individuelles telles que des noms propres (lieux, objets spécifiques, nom de personnes ou d'organismes ».[4]

Les concepts peuvent être reliés à d'autres concepts en attribuant une relation soit hiérarchique ou associative. Ils sont reliés aux termes et ne peuvent pas avoir plusieurs termes préférentiels à la fois juste un seul mais peuvent avoir un nombre infini de termes non préférentiels.

---

[3] CHICHEREAU, Dominique . *Les normes de conception, gestion et maintenance de thésaurus*

[4] DALBIN, Sylvie. 2013. *Livre blanc ISO 25964-1 - Thésaurus pour la recherche documentaire*.

Les termes sont des chaines de caractère, ils peuvent être des mots, des groupes de mots. Ils sont répartis en préférentiels ou non préférentiel et ce sont les préférentiels qui vont renvoyer aux concepts. Toutefois, pour chaque concept il lui faut qu'un seul terme.

Le modèle des données qui manquait aux précédentes normes est défini sous forme d'un diagramme UML. Le modèle de données est décrit dans le livre blanc de l'ISO 25964 comme «un ensemble de classes, chacune représentant un des composants-clés du thésaurus : *Thesaurus, ThesaurusConcept, ThesaurusTerm, ConceptGroup, ThesaurusArray, Note*.
Chaque classe est enrichie d'attributs et de relations avec d'autres classes.»
Cette nouvelle norme permet aussi de se soumettre aux recommandations du W3C parmi lesquelles le format SKOS qui « s'est appuyé dès le démarrage du projet sur les normes existantes, en particulier les normes ISO 2788 or BS 8723-2:2005 .»[5]. Donc, c'est elle qui va accompagner l'élaboration de notre future thésaurus dans lequel il manque la plupart de ces notions présentées ici particulièrement celles qui portent sur les relations sémantiques que nous allons présenter au quatrième chapitre.

[5] DALBIN, Sylvie. 2013. *Livre blanc ISO 25964-1 - Thésaurus pour la recherche documentaire*.

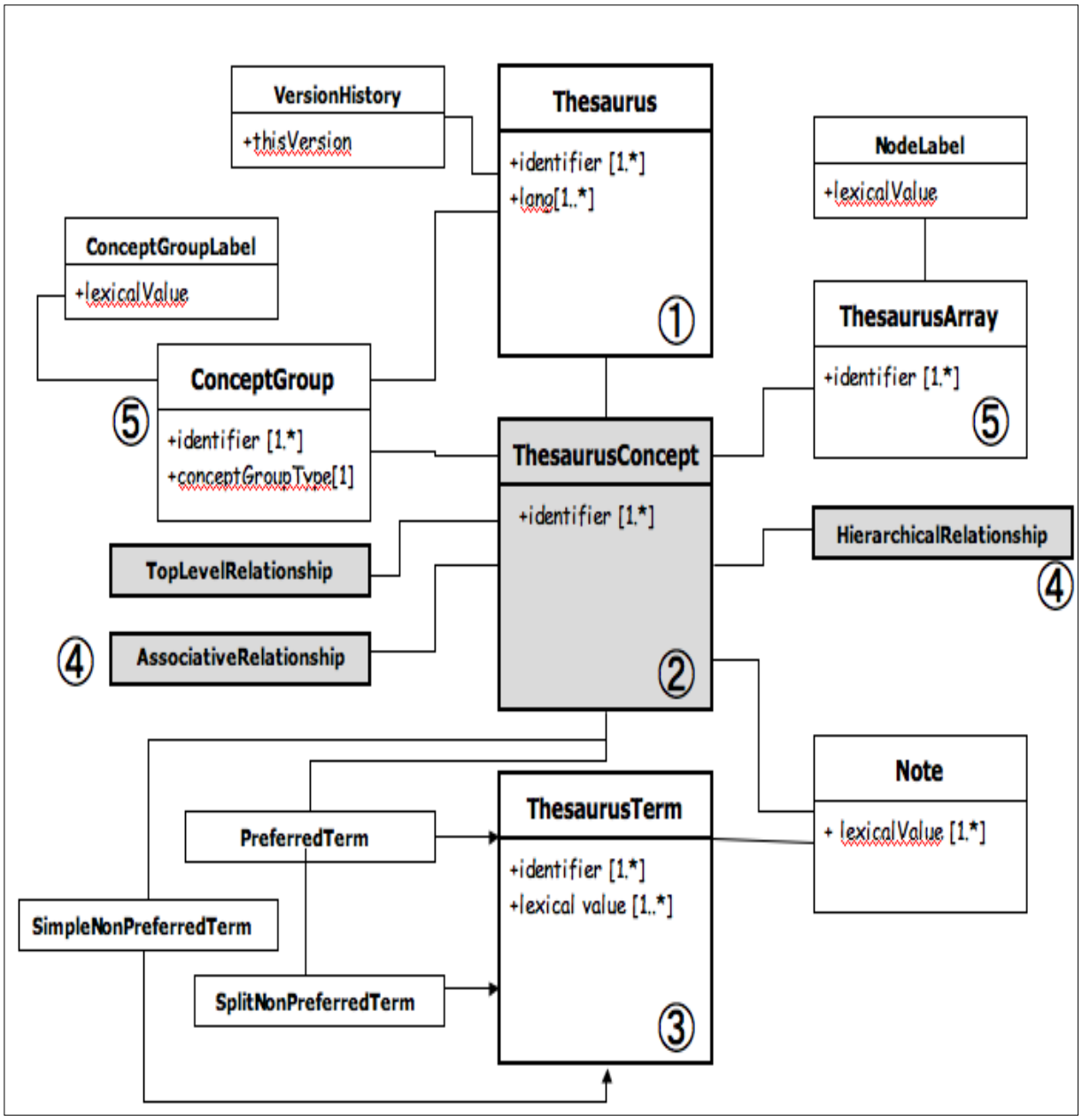

Figure 4 Schéma simplifié du modèle de données ISO 25964-1:2011

2 Projet VocINRA

2.1 Vocabulaire Contrôlé

Dans le domaine des sciences de l'information, un vocabulaire contrôlé est un ensemble de termes spécifiques à un ou plusieurs domaines défini par une organisation pour faciliter la recherche d'information et la représentation de celle-ci. C'est un moyen efficace qui permet de classer les termes afin de lever le voile sur le problème des mots qui sont polysémiques, synonymes, antonymes. Dans un vocabulaire contrôlé, les termes sont rangés suivant une hiérarchie en partant du générique au spécifique. Si nous prenons le mot « fruit », il est le terme générique de « banane », « pomme », « mangue » etc.

Par ailleurs, comme le langage naturel est de nature ambigüe, l'utilisation d'un vocabulaire contrôlé permettra de lever l'ambigüité que présente le langage humain comme le cas de la synonymie des termes ou la polysémie.

La liste des termes qui composent un vocabulaire contrôlé est définie par l'auteur qui choisit les mots en fonction de ses besoins. Dans le cas de VocINRA tous les termes qui sont relatifs à l'agronomie et les domaines associés sont listés ; ils peuvent être en français avec des équivalents d'ans d'autres langues comme l'anglais. Tous les termes n'ont pas la même fonction, nous avons des préférentiels et des non préférentiels. Nous reviendrons largement sur ce point au cours de la rédaction.

2.2 Présentation de VocINRA

VocINRA est le vocabulaire qui regroupe l'ensemble des descripteurs utilisés ou proposés par les documentalistes pour indexer les documents dans ProdINRA, l'archive ouverte institutionnelle de l'INRA.

Certains mots-clés sont empruntés à d'autres sources comme AGROVOC, Termsciences et HAL[6]. VocINRA est composé de mots en français mais des équivalents en anglais ont été donnés à certains termes. Donc il est multilingue.

Le thésaurus VocINRA est un vocabulaire contrôlé qui recouvre les termes du domaine de l'agronomie principalement mais il traite aussi de l'environnement, des sciences exactes, sociales, humaines.

Le thésaurus VocINRA est géré/maintenu grâce à un outil nommé « interface de gestion des mots-clés » (IMC) et qui n'est disponible qu'en interne à l'institut. Tous les mots-clés y sont

[6] http://wiki.inra.fr/wiki/prodinra/Indexation/En+savoir+plus+sur+VOCINRA?xpage=print

stockés. Les mots-clés sont répartis en huit entrée thématique qui porte chacun sur un thème spécifique, d'où leur nom entrée thématique (ET) avec chacun un numéro (de 1 à 8).

Entrée thématique	Description - Exemples
ET1 - Objet d'études	- Organismes ou systèmes biologiques : les organismes vivants et/ou leurs différents niveaux anatomiques (organes, systèmes organiques ou éléments cellulaires). Cela couvre l'ensemble du monde vivant, des micro-organismes jusqu'aux organismes supérieurs. Les organismes sont nommés par leurs noms communs et latins. Exemples : *chêne, cellule bactérienne, cépage apyrène, nerf gastrique, quercus petraea, racine*
ET2 Question sociétale et finalité, contexte	La problématique générale de la recherche, la visée de l'étude sur la base d'enjeux scientifiques et sociétaux par rapport auxquels elle est positionnée. Exemples : *agriculture de Montagne, alimentation animale, changement climatique, condition de fermentation, projet de développement, relation plante-insecte, variabilité génétique*.
ET3 Démarche, discipline	Référentiel des disciplines : *Agronomie, Architecture, Aménagement de l'espace, Biologie cellulaire, Toxicologie et chaîne alimentaire, etc.*
ET4 Échelle d'étude	Les dimensions du système étudié, que celui-ci représente une population ou une colonie d'individus dans un écosystème naturel ou artificiel, un système réduit à l'individu, un organe, un système d'organes ou bien des éléments cellulaires. Les termes employés ici ne reprennent pas ceux de l'entrée thématique "organismes étudiés" mais en sont leur caractéristique en termes d'échelle. Exemples: *alvéole, champ cultivé, chênaie, colonie d'abeilles, forêt artificielle, organe reproducteur végétal, parcelle, région viticole, tube digestif.*

ET5 Localisation géographique	La localisation géographique de l'étude. Exemples : *Vaucluse, PACA, Camargue, forêt landaise, lac Léman.*
ET6 Dispositif technique et méthode d'étude	Les moyens techniques, les plates-formes, les outils et les méthodologies mis en œuvre lors des phases d'expérimentation et de traitement des données. Exemples : *accélérateur de particules, élongation du pollen, méthode stochastique, microscopie électronique, télédétection, transfert embryonnaire.*
ET7 Composé chimique, facteur du milieu	Les composés ou facteurs physico-chimiques, biotiques ou abiotiques présents sous forme atomique ou moléculaire. Cela couvre les composés biochimiques présents dans les organismes vivants aussi bien que les facteurs physico-chimiques du milieu extérieur. Exemples : *1,2,4-triazole-3-ylamine, acide nucléique, acide aminé, hormone, lipide du lait, polluant.*
ET8 Phénomène, processus et fonction	Les mécanismes biologiques ou physico-chimiques fondamentaux qui déterminent les processus, fonctions ou propriétés d'un organisme, d'un écosystème ou d'un milieu abiotique. Exemples :*absorption de protéines, conduction nerveuse, division cellulaire, insuffisance pancréatique, photosynthèse, oxydation, propriété organoleptique.*

Figure 5 Entrées Thématiques composant VocINRA avec leur description

Source : Url: http://wiki.inra.fr/wiki/ProdINRA/Indexation/En+savoir+plus+sur+VOCINRA

Dans l'interface des mots-clés, les termes ont chacun un statut : *valide, invalide et attente validation.* Pour chaque terme qu'il soit valide, invalide ou en attente un rôle lui est attribué : *principal, associe, traduit.*

Les mots-clés entretiennent des relations entre eux : *générique, associé (*autour d'une même thématique*), traduit* (équivalent d'un terme dans une autre langue).

Entrées thématiques (ET)	Valide	Invalide	Attente Validation	Total
ET1	30676	514	1505	32695
ET2	467	571	57	1095
ET3	743	10	0	753
ET4	3705	327	66	4098
ET5	1321	35	70	1426
ET6	18934	223	515	19672
ET7	17436	177	262	17875
ET8	16593	188	405	17186
Total	89875	2045	2880	94800

Figure 6 taille du vocabulaire exporté de l'IMC

2.3 Exemples de Vocabulaire agronomique

2.3.1 AGROVOC

AGROVOC tout comme VocINRA est un vocabulaire contrôlé. Il a été conçu pour rassembler les termes qui sont en rapport avec les domaines d'activités de la FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture) à savoir « l'alimentation, la nutrition, l'agriculture, la pêche, la foresterie, l'environnement »[7]. C'est un thésaurus

[7] http://aims.fao.org/fr/AGROVOC

multilingue avec une représentation des concepts dans 23 langues. Le nombre de termes préférentiels et non préférentiels en français est de 38572.[8] Il est ouvert à tout le monde car il est partagé sur le web grâce à l'outil VocBench. C'est le même logiciel que nous avons décidé de tester pour voir s'il pourra gérer notre vocabulaire. Il offre la possibilité d'y apporter des modifications en faisant des propositions de suppression ou d'ajout de termes entre autres. C'est un thésaurus qui répond aux besoins du Web sémantique car il est défini sous un schéma de concept Skos-XL .

AGROVOC s'est aligné avec d'autres thésaurus dans le domaine de l'agriculture, donc des techniques ont été développées à cet effet et nous essayerons de les reprendre pour pouvoir s'aligner à notre tour avec ce vocabulaire et GACS que nous présenterons dans le chapitre suivant.

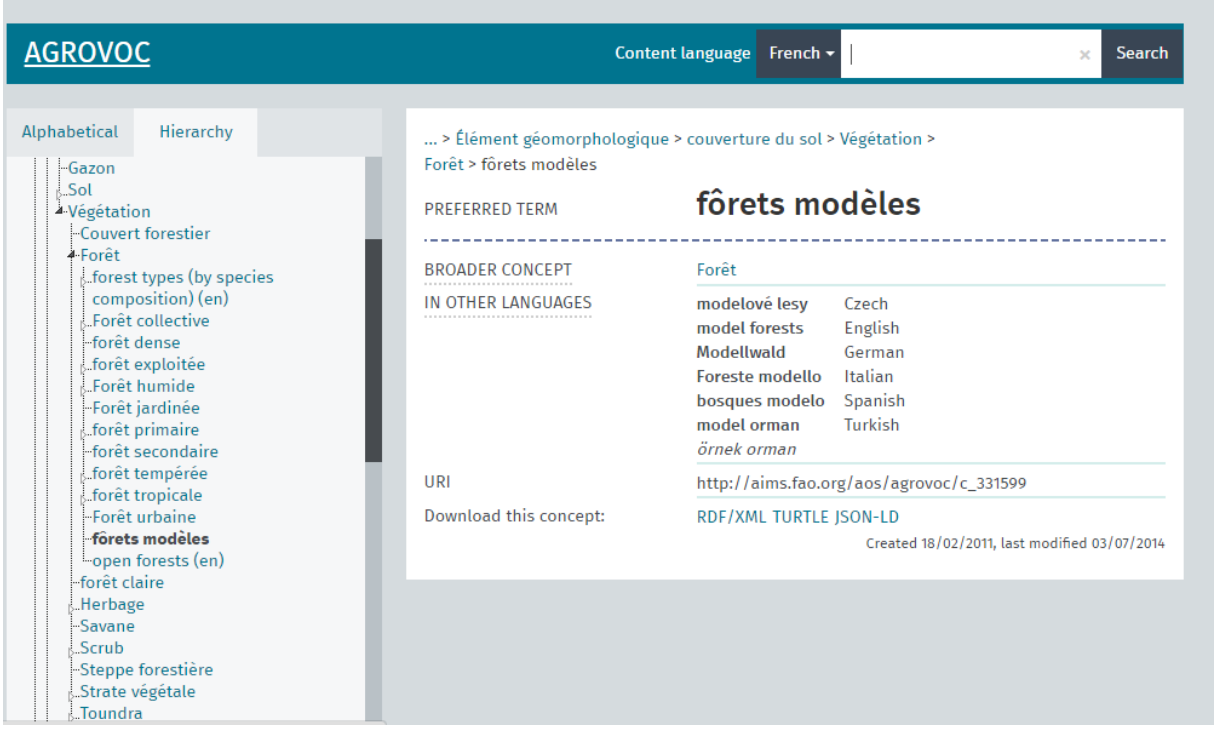

Figure 7 capture d'écran d'un extrait du vocabulaire d'AGROVOC

URL de l'image : **http://oek1.fao.org/Skosmos/AGROVOC/en/page/c_331599?clang=fr**

[8] http://aims.fao.org/standards/AGROVOC/concept-scheme

2.3.2 GACS

GACS[9] (Global Agriculture Concept Scheme) est un thésaurus dont la première version est publiée en Beta. Il est issu de la collaboration entre la FAO (AGROVOC), CABI (CAB thesaurus) et NAL (NAL thesaurus) pour regrouper autour d'un schéma de concept Skos tous les concepts que ces trois vocabulaires ont en commun comme ils opèrent tous dans le domaine de l'agriculture. Chacun des thésaurus a fourni les 10000 concepts[10] les plus utilisés dans leur vocabulaire. Tout comme AGROVOC, la version beta de GACS est éditée en projet dans l'outil VocBench.

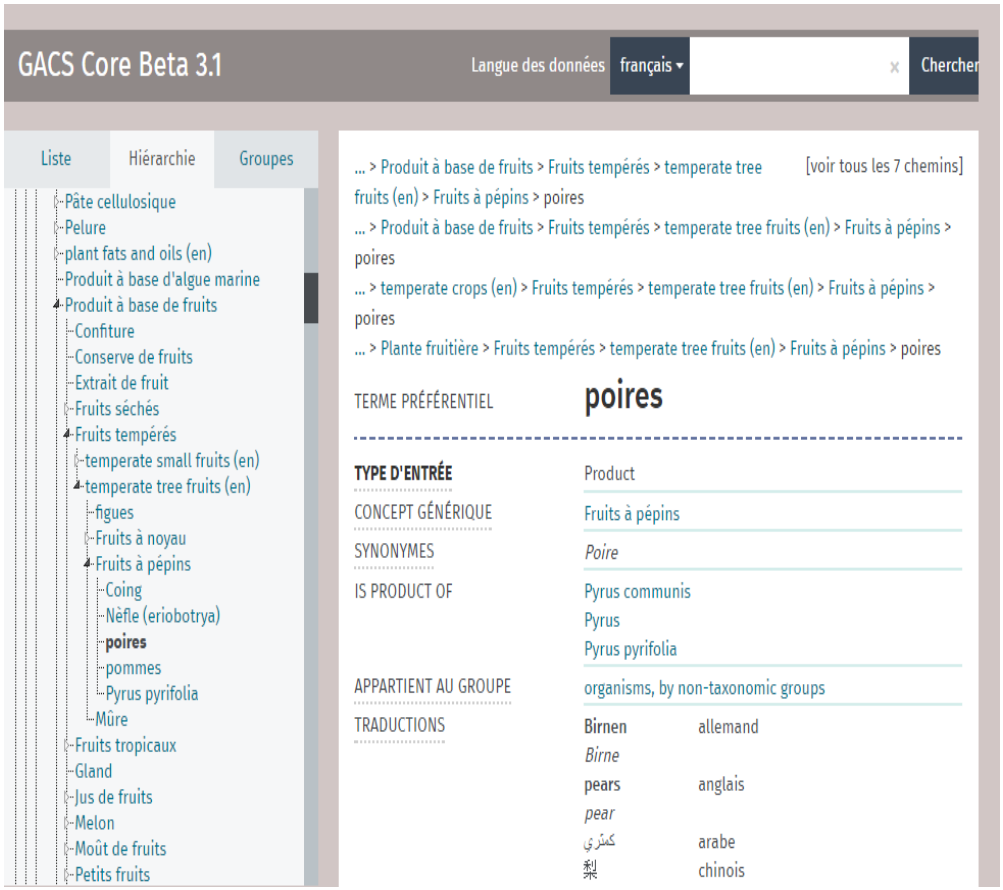

Figure 8 capture d'écran d'un extrait de GACS

URL de l'image: **http://tester-os-kktest.lib.helsinki.fi/gacsdemo/gacs/fr/page/C3071**

[9] http://aims.fao.org/activity/blog/global-agricultural-concept-scheme-gacs-collaborative-integration-three-thesauri

[10] http://ist.blogs.inra.fr/technologies/2015/06/02/global-agricultural-concept-scheme-gacs/

2.4 Outils de Travail

2.4.1 IMC

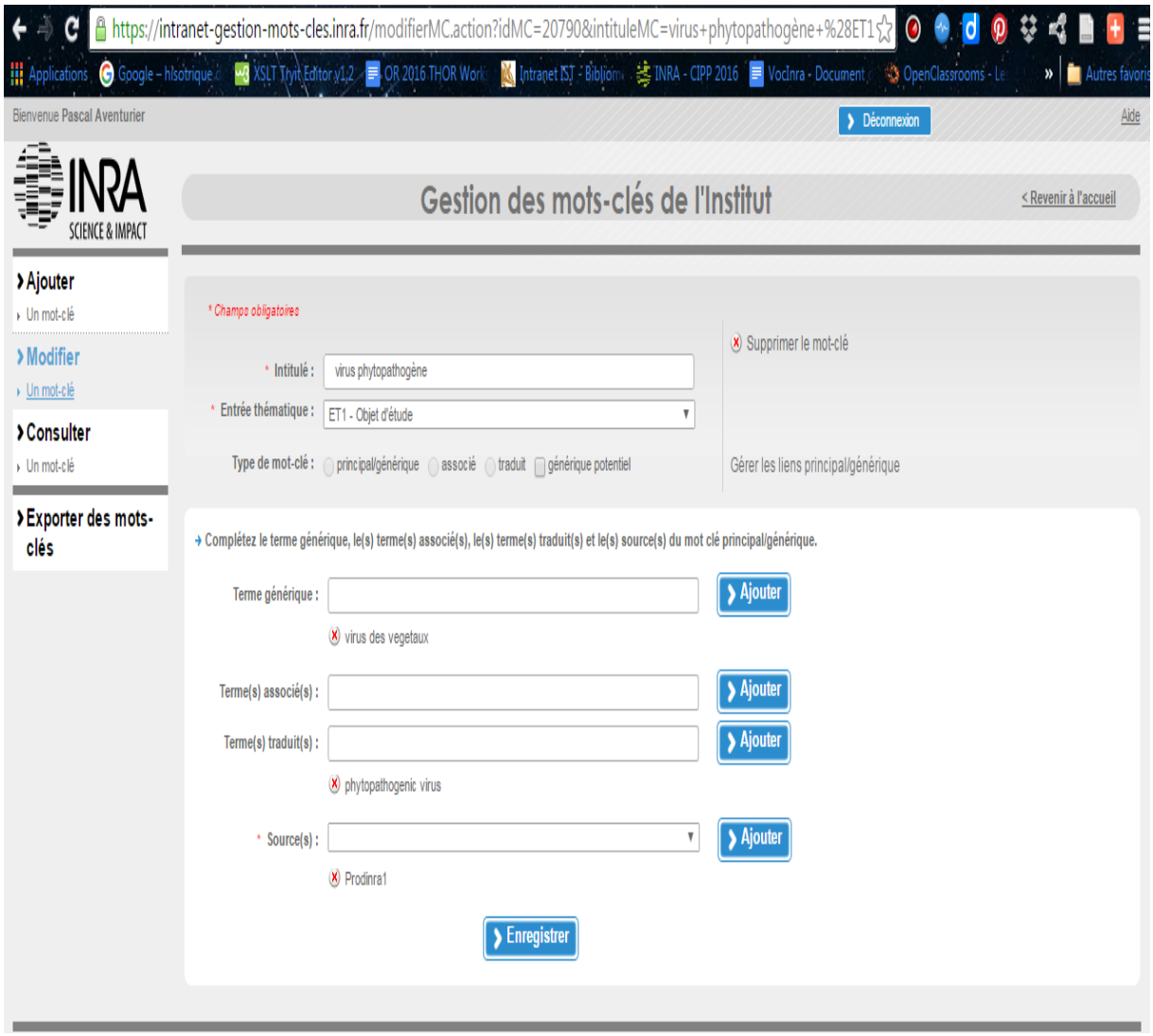

Figure 9 capture d'écran de l'interface de l'IMC

L'IMC est l'interface de gestion des mots-clés qui est développé en interne. C'est dans cet outil de gestion que les mots qui composent notre vocabulaire sont stockés. Il gère le thésaurus de l'INRA qui est réparti en huit micro thésaurus et exportable sous plusieurs format comme le XML, le CSV, l'XLS ou le PDF. Le vocabulaire est managé de façon séparée avec les huit entrées thématiques qui ne sont pas réunies.

A partir de l'IMC, un utilisateur peut proposer un mot clé qui sera validé ou bien rejeté. Il peut aussi modifier un mot-clé en apportant quelques corrections si des fautes ont été

commises au moment de la saisie. Pour déplacer un terme d'une entrée à une autre, il faudra le supprimer d'abord pour pouvoir le recréer dans l'entrée thématique ce qui est vraiment fastidieux. Le référentiel ne permet pas non plus de relier un terme spécifique à plusieurs termes génériques. Par exemple, le terme «Espagne » ne peut pas avoir pour termes génériques « Pays Méditerranéens » et « Europe ». Seul le choix entre un des deux est possible. Il ne répond pas aux besoins du web sémantique ni à certaines des règles établies par l'iso 25964 à savoir les relations sémantiques qui y sont pauvres. Trois relations seulement sont définies : générique, associé et traduit.

Les fonctionnalités que l'IMC propose sont trop limitées et ne satisfont pas aux demandes du Web sémantique : le format actuel (XML et autres) n'est pas celui proposé par le W3C (SKOS). Il n'est pas conçu malheureusement pour supporter ce format donc le remplacer par un outil qui répond à ces recommandations serait envisageable, d'où la mise en place de VocBench.

2.4.2 VOCBENCH

VocBench[11] a été conçu à travers une collaboration entre l'Organisation pour l'alimentation et l'agriculture (FAO) des Nations Unies et l'Université de Rome Tor Vergata[12] pour gérer le vocabulaire de la FAO (AGROVOC).

Cet outil est développé pour permettre aux thésaurus d'être en phase avec les recommandations du Web sémantique. Il ne prend en charge que les fichiers au format recommandé pour les thésaurus à savoir le SKOS. VocBench permet de gérer des vocabulaires monolingues ou multilingues. Plusieurs fonctionnalités y sont disponibles. Il est possible de supprimer un terme ou un concept ou d'en créer un nouveau. Dans VocBench, un utilisateur peut modifier l'organisation des concepts en déplaçant un concept grâce l'onglet « *move* concept» pour en faire par exemple un *narrower* d'un autre concept ou lier concept *narrower* à deux concepts génériques. Dans l'IMC, ce n'était pas possible de lier un terme à deux termes génériques. La relation générique était le plus haut niveau de la hiérarchie dans l'IMC, or dans VocBench le plus haut niveau hiérarchique est représenté à travers les tops concepts.

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

[12] http://aims.fao.org/fr/vest-registry/tools/vocbench

Cet outil dispose aussi d'une interface SPARQL qui permet d'interroger son vocabulaire (SPARQL 1.1 Query) ou d'y apporter quelques modifications (SPARQL Update). Ce qui est bien car l'onglet « *move concept* » ne peut déplacer les concepts qu'un par un. Toutefois les modifications apportées dans VocBench ne sont pas effectives, elles sont à valider par un administrateur car l'outil est développé sous forme de gestion de projet.

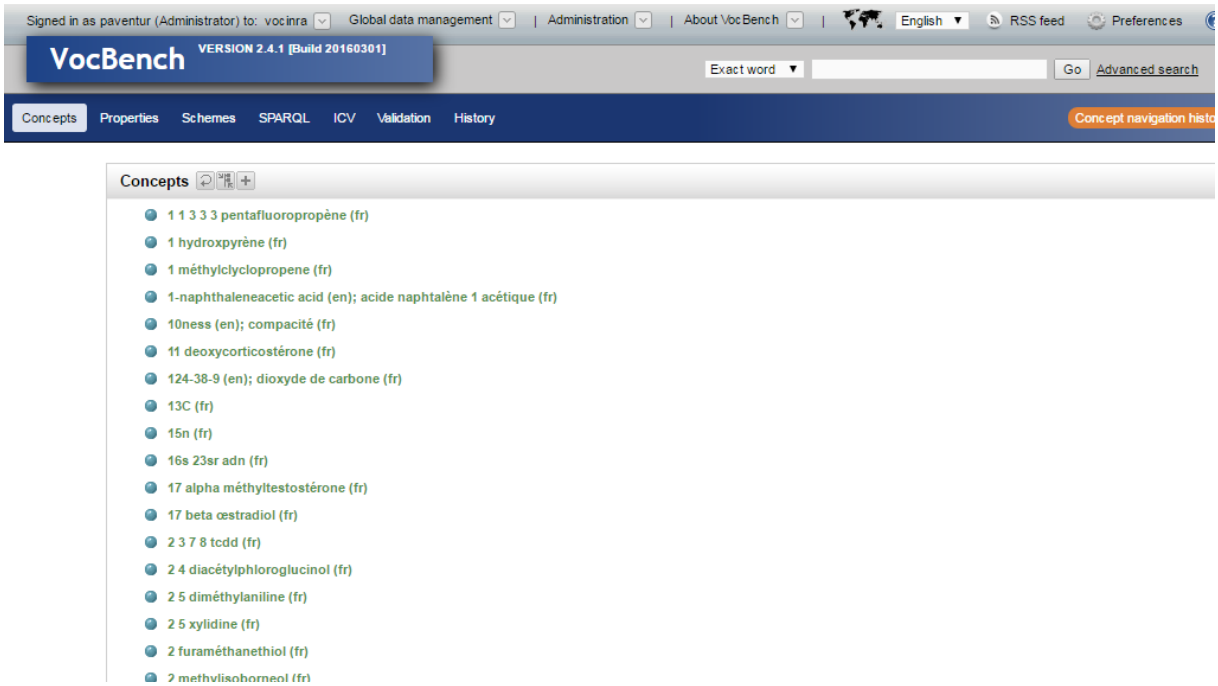

Figure 10 capture d'écran d'un extrait du vocabulaire chargé dans VocBench

Url : http://w3.avignon.inra.fr/VocBench-2.4/#Concepts

2.4.3 OPENREFINE

Anciennement apparu sous le nom de Google Refine, OpenRefine[13] est un logiciel créé par Google pour manipuler, nettoyer des données en masse et aussi les transformer. Un livre intitulé *Using OpenRefine* et écrit par Ruben Verborgh et Max De Wilde explique en détail les fonctionnalités de OpenRefine et comment s'en servir. Nous nous sommes le plus basé sur le wiki[14] rédigé par Sophie Aubin pour mieux aborder OpenRefine. C'est un logiciel à installer sur son poste de travail, il dispose d'une interface web où les manipulations pourront

[13] http://openrefine.org/
[14] http://wiki.inra.fr/wiki/traitementsdocumentaires/Main/OpenRefine

être faites. OpenRefine accepte différents formats d'import à savoir le TSV, CSV, LE XLS, XLSX, XML, RDF/XML, RDF N3 TRIPLES. L'export aussi peut se faire avec ces formats, d'autres formats sont aussi offerts comme le HTML, l'ODT, le templating, etc. Pour exporter ses données en RDF, l'installation de l'extension RDF dans l'espace de travail d'OpenRefine est nécessaire. Il dispose également de plusieurs fonctionnalités toutes importantes dont nous avons exploité après la transformation de nos données en SKOS que nous verrons dans le troisième chapitre.

Nous allons citer toutefois quelques fonctionnalités comme l'onglet « text facet » qui nous a été très utile car il permet de visualiser les données d'une colonne et aussi de les compter. Pour apporter quelques modifications à ces fichiers l'onglet « edit cell » pourrait servir. Dans l'onglet « edit cell » il y a le « transform » qui se sert du langage grell pour permettre de modifier une valeur de donnée, le « common transform » agit sur toutes les données des cellules d'une colonne. Pour mettre ses données en majuscules ou en minuscules le « common transform » permet de le faire. Nous nous sommes servis de cet onglet pour mettre en majuscules les données de l'entrée thématique 5 composées d'entités nommées.

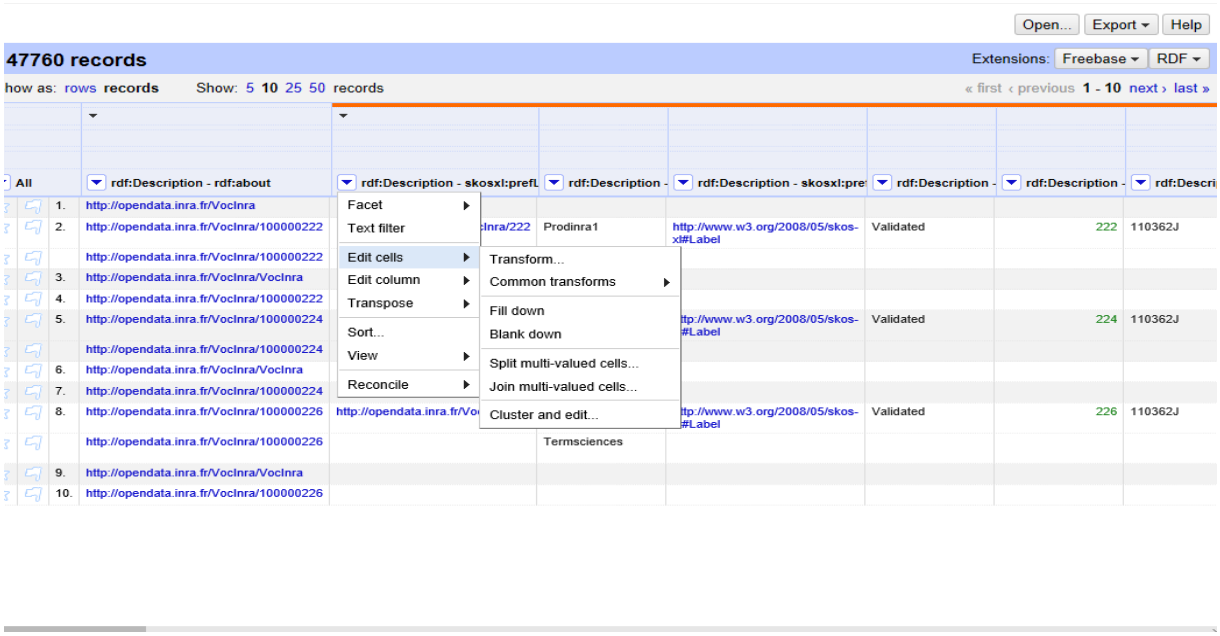

Figure 11 capture d'écran d'un projet OpenRefine

3 Corpus de travail

3.1 Corpus IMC

Notre premier corpus de travail extrait de l'IMC est constitué de huit fichiers au format XML. Ces fichiers, nommés « entrée thématique », forment le référentiel des mots-clés de l'INRA composé des descripteurs/termes qui sont validés, proposés ou rejetés.

N°	Entrée thématique	fichier
1	Objet d'étude	MC_ET_1.xml
2	Question sociétale et finalité, contexte	MC_ET_2.xml
3	Démarche, discipline	MC_ET_3.xml
4	Echelle d'étude	MC_ET_4.xml
5	Localisation géographique	MC_ET_5.xml
6	Dispositif technique et méthode d'étude	MC_ET_6.xml
7	Composé chimique, Facteur du milieu	MC_ET_7.xml

Figure 12 entrées thématiques exportées au format XML

Dans nos fichiers XML extraits de l'IMC (outil interne de gestion des mots-clés) les termes sont organisés en quatre catégories

- Mots-clés Principaux
- Termes Génériques
- Termes Associes
- Termes Traduits

Cette organisation définit le statut des mots-clés (Principal, traduit et associé) et la relation qui existent entre eux (générique, associé traduit).

Nous avons une liste de mots-clés principaux qui regroupe l'ensemble des termes qui sont spécifiques aux termes génériques. Dans le fichier XML, l'arbre auquel appartient un terme spécifique, c'est-à-dire son terme générique, est défini.

```
<libelle>cybernétique</libelle>
<motClePrincipalList xsi:type="motClePrincipal"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
<codeEtat>1</codeEtat>
<libelle>VALIDE</libelle>
</etat>
<langue>
<idFonctionnel>1</idFonctionnel>
<libelle>FRANCAIS</libelle>
</langue>
<libelle>robotique</libelle>
<arbre>
<idFonctionnel>22419</idFonctionnel>
</arbre>
```

Dans cet extrait, le mot *cybernétique* est relié au terme *robotique* qui est son générique.
Les termes génériques représentent les termes qui sont les éléments pères de certains termes. Ils sont reliés à ces derniers car ils sont considérés comme ceux ayant une définition plus vaste. Dans les fichiers de l'IMC, les termes ne peuvent pas avoir plusieurs termes génériques à la fois. Donc un terme générique peut être relié à n mots-clés principaux mais aucun mot-clé principal n'a la possibilité d'appartenir à plusieurs termes génériques. Cette fonctionnalité n'est pas disponible dans l'IMC
Les termes associés représentent les termes synonymes ou quasi synonymes des termes principaux et permettent aux utilisateurs de disposer d'un champ de recherche plus large. Comme le thésaurus est multilingue, des équivalents en anglais ou en espagnol ont été créés pour certains termes, d'où la liste des termes traduits.

Chaque terme a un identifiant qui lui est propre, une date de création, un auteur, l'entrée thématique à laquelle elle appartient est aussi définie.
Les identifiants des termes sont compris dans la balise « *idFonctionnel* » et vont de 1 à 94800 (nombre total des mots-clés). La langue de chaque mot –clé est identifié (1 pour le français, 2 pour l'anglais et 3 pour l'espagnol). Les entrées thématiques sont aussi numérotées de 0 à 8 dans le fichier XML. L'identifiant 0 est une anomalie, nous y reviendrons.

```
<idFonctionnel>34528</idFonctionnel>
<identifiantEntreeThematique>4</identifiantEntreeThematique>
<langue>
<idFonctionnel>2</idFonctionnel>
<libelle>ANGLAIS</libelle>
</langue>
<libelle>holocene epoch</libelle>
</motCleComplet>
```

L'auteur de chaque mot clé est répertorié avec un matricule ainsi que la date de création du mot. Si un terme a été supprimé ou déplacé, il devient inactif et sa date d'inactivation va figurer dans le fichier XML.

```
<motCleComplet>
<auteurCreation>
<matricule>110362J</matricule>
</auteurCreation>
```

Pour rappel, les mots ont chacun un état ou statut : ils sont actifs quand ils sont **validés**, en **attente de validation** quand la proposition n'est pas encore validée et **inactifs** quand ils sont supprimés définitivement du vocabulaire ou supprimés pour être recréés dans une autre entrée thématique.

```
<motCleComplet>
  <auteurCreation>
    <matricule>15816R</matricule>
  </auteurCreation>
<candidatPere>false</candidatPere>
<categorieMC>PRINCIPAL</categorieMC>
<etat>
```

<codeEtat>2</codeEtat>
<libelle>ATTENTE_VALIDATION </libelle>
</etat>
<idFonctionnel>99032</idFonctionnel>
<identifiantEntreeThematique>0</identifiantEntreeThematique>
<libelle>unité cartographique de sol</libelle>
<source>
<idFonctionnel>10</idFonctionnel>
<libelle>ProdINRASource</libelle>
<url>non-determiné</url>
</source>
</motCleComplet>

<motCleComplet>
<auteurCreation>
<matricule>110362J</matricule>
</auteurCreation>
<candidatPere>false</candidatPere>
<categorieMC>ASSOCIE</categorieMC>
<etat>
<codeEtat>3</codeEtat>
<libelle>INVALIDE </libelle>
</etat>
<idFonctionnel>35195</idFonctionnel>
<identifiantEntreeThematique>0</identifiantEntreeThematique>
<langue>
<idFonctionnel>1</idFonctionnel>
<libelle>FRANCAIS</libelle>
</langue>
<libelle>limon</libelle>
</motCleComplet>

3.2 Corpus ProdINRA

Notre deuxième corpus est un export de toutes les indexations réalisées dans l'archive ProdINRA. Sur 225067 références de documents seulement 62% étaient indexés. Ces fichiers sont au format XML va être utilisé pour le nettoyage de notre corpus IMC. En effet, le corpus IMC comporte un nombre important de mots-clés qui sont pour la plupart issus de l'indexation de ProdINRA et tous ces mots n'ont plus la même la valeur. Il y en a qui ne sont plus du tout utilisés pour indexer les documents. Donc, l'idéal serait de retrouver tous les termes qui ne sont utilisés qu'une seule fois durant ces vingt dernières années afin de les éliminer du vocabulaire et réduire sa taille. Les termes à supprimer vont être extraits du corpus ProdINRA avec leur identifiant pour obtenir leur fréquence. Dans ce corpus, ce sont les balises qui vont nous intéresser sont les balises « *thematic* » puisque les termes et les identifiants à extraire s'y trouvent.

```
<thematic>
<identifier>1</identifier>
<name>Objet d'étude</name>
<inraClassification>
<inraClassificationIdentifier>1458</inraClassificationIdentifier>
<usedTerm>arabidopsis thaliana</usedTerm>
<genericTerm>arabidopsis</genericTerm>
</inraClassification>
<inraClassification>
<inraClassificationIdentifier>17497</inraClassificationIdentifier>
 <usedTerm>rosette</usedTerm>
  </inraClassification>
  <inraClassification>
<inraClassificationIdentifier>18859</inraClassificationIdentifier>
<usedTerm>surface foliaire</usedTerm>
<engTerm>leaf area</engTerm>
</inraClassification>
        </thematic>
```

3.3 Filtrage des données

Dans cette partie nous avons adopté une méthode en Traitement automatique des langues qu'est l'extraction d'information qui peut se faire à l'aide d'outils déjà développés ou de programmes informatiques. L'information à extraire ici est l'ensemble des mots-clés issus de l'indexation et leur fréquence pour pouvoir supprimer les termes qui n'apparaissent qu'une seule fois. Ce filtrage est réalisé sur le deuxième corpus extrait de ProdINRA. Étant donné que le calcul ne pouvait pas se faire manuellement à cause du volume important de fichiers, nous nous sommes servis d'un programme en java que nous avons développé à cet effet.

Le programme écrit débute par une lecture pour parcourir le fichier XML en ne tenant compte que des balises « thematic » dans lesquelles se trouvent les parties qui nous intéressent à savoir les balises « usedTerm » et « InraclassificationIdentifier ». Donc, s'il trouve un terme il va continuer à parcourir le fichier pour pouvoir calculer sa fréquence. Ainsi nous avons pu extraire les termes avec leur identifiant et le nombre de fois qu'ils apparaissent. **57454** termes furent repérés et chargés dans OpenRefine (voir section 2). Ce chargement nous a aidés à repérer les termes avec une occurrence égale à **1** et de les filtrer grâce à l'onglet « text facet ».

Le calcul des nombres d'occurrences, réalisé avec le programme Java, peut aussi se faire directement avec OpenRefine. Nous l'avons aussi testé. Il faut importer les fichiers XML dans OpenRefine et sélectionner la balise « thematic » dans laquelle nous avons, les termes à compter, leur identifiant, leur langue. Une fois les fichiers chargés, il faut fusionner les colonnes contenant les termes et les identifiants, et ensuite faire une facette avec la nouvelle colonne obtenue. Cette facette affichera la fréquence de chaque terme.

termes	identifiants	fréquence
france	40402	12055
europe	39882	6615
modélisation	43833	6211
bovin	2683	5751
protéine	70466	5682
arbre forestier	1502	5400
croissance	38156	4755
poisson	15191	4626
sol	73244	4491
ovin	13669	4409
plante légumièr	14962	4191
porcin	15486	4172
vitis vinifera	20862	4134
vigne	20479	3765

Figure 13 extrait du résultat de la facette d'OpenRefine

Quand nous avons chargé le résultat de notre programme d'extraction, nous avons remarqué que certains termes avaient pour identifiant **0** et d'autres termes qui étaient écrits de la même manière avaient des identifiants différents**.** Le problème provient de l'export qui nous a été fourni. Par ailleurs, nous avons essayé de le régler en chargeant tous les termes sans les identifiants dans OpenRefine pour vérifier s'il n'y avait pas de doublons. En fait, le programme comme il calcule la fréquence des termes avec leur identifiant, il ne tiendra pas compte des termes qui sont semblables. Donc, si un terme a un mauvais identifiant, il sera compté de manière isolée sans tenir compte des termes semblables avec le bon identifiant.

La facette dans OpenRefine nous a permis de retrouver 347 termes sur les 20709 qui étaient récupérés avec une occurrence égale à 1.

Ensuite, avec le corpus ProdINRA nous avons pu repérer tous les termes qu'il fallait exclure de notre vocabulaire. Le nettoyage s'est fait avec le premier corpus qui constitue notre thésaurus.

Une fois les termes repérés, il fallait les supprimer. Pour la suppression des termes un autre programme j'ava a été développé. Dans un premier temps, le programme lit le fichier XML de départ c'est-à-dire celui que nous avons extrait de l'IMC (notre premier corpus) et le parcourt pour trouver les termes stockés dans une liste qu'on lui a fournit au départ. Ensuite, il les supprime un par un pour régénérer à la fin un nouveau fichier XML.

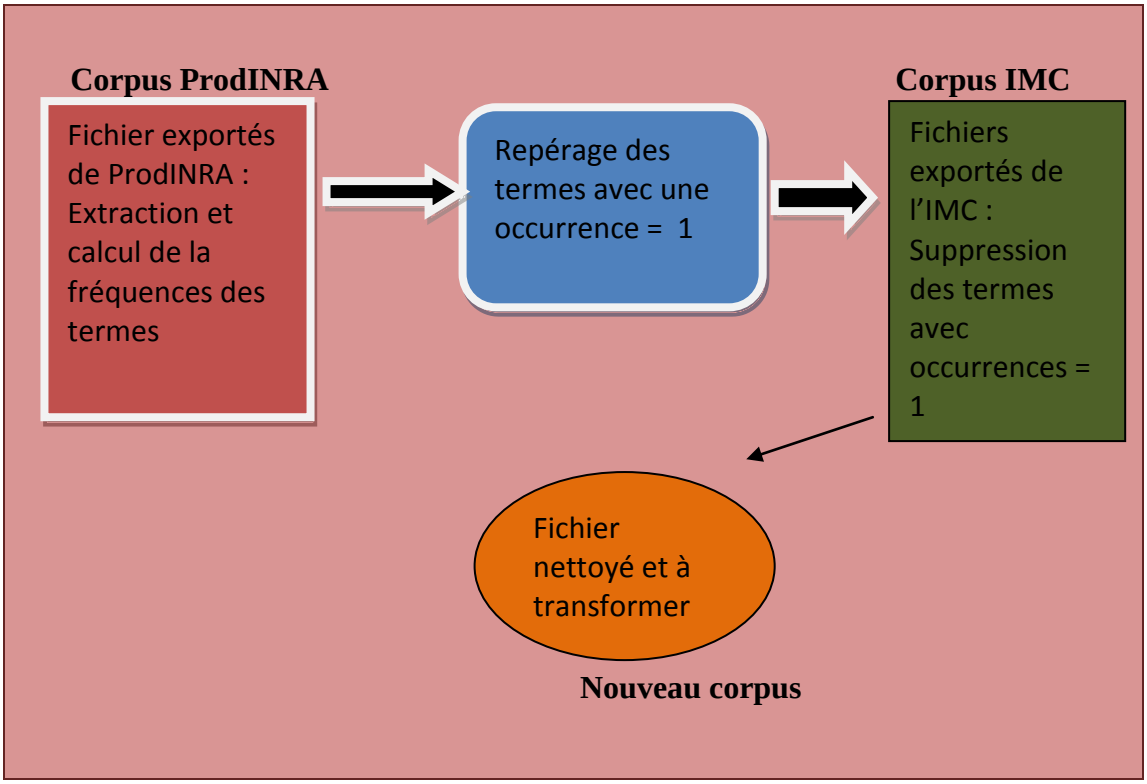

Figure 14 tableau récapitulatif des étapes du filtrage

3.4 Transformation des données en SKOS

L'évolution du web repose sur des recommandations mises en place par le W3C à travers le web sémantique. Parmi ces recommandations, il y a le standard SKOS, « un modèle de données partagé pour échanger et relier des systèmes d'organisation de connaissances sur le Web »[15]. Il est destiné aux thesauri, aux taxonomies, aux classifications ou aux systèmes de vedettes matières. SKOS se sert des triplets RDF pour exprimes ses données donc n'importe quelle syntaxe RDF peut être employée.

Par ailleurs le logiciel VocBench choisi par l'équipe de la DIST pour la future gestion et exploitation de notre thésaurus est développé dans le but d'être en phase avec les besoins du web sémantique. VocBench ne prend en entrée que les fichiers SKOS ou SKOS-XL (voir section X pour la description de ces formats). Un des objectifs du stage était de changer notre format de départ qui du XML vers du Skos pour être en phase avec les recommandations du W3C dans le cadre de l'élaboration des thésaurus qui seront interopérables à travers le web de données.

Pour la transformation , une première version de feuille de style XSL a été réalisée par Sophie Aubin du pôle Geco de l'INRA. Lors de ce stage nous avons testé cette feuille, et corrigé ce qui générait une structure erronée. Dans cette feuille de style, les quatre catégories présentes dans nos fichiers XML vont être associées à quatre templates:

- La catégorie <termeGenerique> va devenir les « broader »
- La catégorie <motClePrincipalList> regroupe les « narrower »

Les termes issus de ces deux catégories sont tous des termes préférentiels (« Skos: prefLabel»)

- La catégorie <termeAssocieList> sera la classe « Skosxl:altLabel »
- La catégorie <termeTraduitList> sera divisée en deux classe : celle des termes non-préférentiels « Skosxl:altLabel » et des termes préférentiels « Skos:prefLabel ».

Dans le fichier XML d'origine, il n'y avait pas une organisation en concepts : tout était termes. Donc, il fallait passer d'une organisation en termes vers une organisation en concepts ce qui n'était pas facile car il fallait retrouver les termes des différentes catégories qui forment un réseau entre eux pour en faire des Skos : Concept.

Nous pouvons citer en exemple *:* le concept *protéine de liaison (fr)* qui regroupe les termes *:*
protéine de liaison (fr) <skos:prefLabel xml:lang="fr">protéine de liaison</skos:prefLabel>
binding proteins (en); <skos:altLabel xml:lang="en">binding proteins</skos:altLabel>
transport proteins (en) <skos:altLabel xml:lang="en">transport proteins</skos:altLabel>

[15] http://www.sparna.fr/Skos/SKOS-traduction-francais.html#L1045

carrier proteins(en) <skos:altLabel xml:lang="en">carrier proteins</skos:altLabel>

Ces concepts sont définis en ajoutant le préfixe 100000 à l'idFonctionnel d'un terme principal.

Pour se faire, l'identification des termes va changer car le préfixe 100000 sera ajouté à l' « *idFonctionnel* » du terme principal.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix skosxl: <http://www.w3.org/2008/05/skos-xl#> .
@prefix dc11: <http://purl.org/dc/elements/1.1/> .
@prefix ns0: <http://art.uniroma2.it/ontologies/VocBench#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dc: <http://purl.org/dc/terms/> .

<http://opendata.inra.fr/VocINRA> a skos:ConceptScheme .
<http://opendata.inra.fr/VocINRA/70540>
 a skosxl:Label ;
 dc11:source "ProdINRA1" ;
 ns0:hasStatus "Validated" ;
 skos:notation "70540"^^xsd:int ;
 dc:creator "110362J" ;
 skosxl:literalForm "protéine de liaison"@fr .

<http://opendata.inra.fr/VocINRA/10000070540>
 skosxl:prefLabel <http://opendata.inra.fr/VocINRA/70540> ;
 skos:prefLabel "protéine de liaison"@fr ;
 skos:notation "10000070540"^^xsd:int ;
 skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 ns0:hasStatus "Validated" ;
 skos:scopeNote "Composé chimique ou facteur milieu" .
```

Les différents statuts, **Valide**, **Attente Validation** et **Invalide**, vont être appelés différemment se conformer à VocBench. Nous aurons alors les statuts « Validated », « Proposed » et « Deprecated » qui vont être remplacés respectivement par **Valide**, **Attente Validation** et **Invalide.** Les entrées thématiques et les langues sont toujours présentes mais sans leur identifiant. Les entrées thématiques seront dans la balise « Skos:scopeNote » et la langue dans

les balises « Skos:prefLabel » et « Skosxl:altLabel » précédés du libellé « xml:lang= » qui est une manière standard d'indiquer la langue en XML RDF.

La structure d'un thésaurus repose sur une hiérarchie, du plus générique au plus spécifique, mais tous les génériques n'ont pas le même statut dans la hiérarchie. Il faudra désigner ce qui seront au plus haut niveau, i.e. les « Skos:topConcept ». Dans notre feuille de style, un test avait été mis en place pour identifier les termes qui n'ont pas de générique et d'en faire des tops concepts. Malheureusement, le test a fonctionné mais pas que sur ceux qui n'ont pas de terme générique. Il a pris aussi tous les termes génériques et les a transformés en top concept. Après de nombreuses tentatives de correction, nous avons réalisé que la représentation des données du fichier source ne permettait pas de réaliser correctement ce test. Nous avons réussi à résoudre ce problème grâce à OpenRefine. Nous avons fait une facette sur la colonne des « narrower » et celle des « top concept ». Tous les termes qui apparaissent deux fois sont à éliminer de la liste des « top concept ».

Par ailleurs, dans VocBench il faut un schème (Skos :ConceptScheme) qui permettra d'afficher les concepts et une URI de base qu'il faudra renseigner au moment du chargement, ils ont aussi été créés.

En résumé, tous les éléments du fichier XML ont été repris dans notre fichier RDF même si pour certains, la notation a été modifiée.

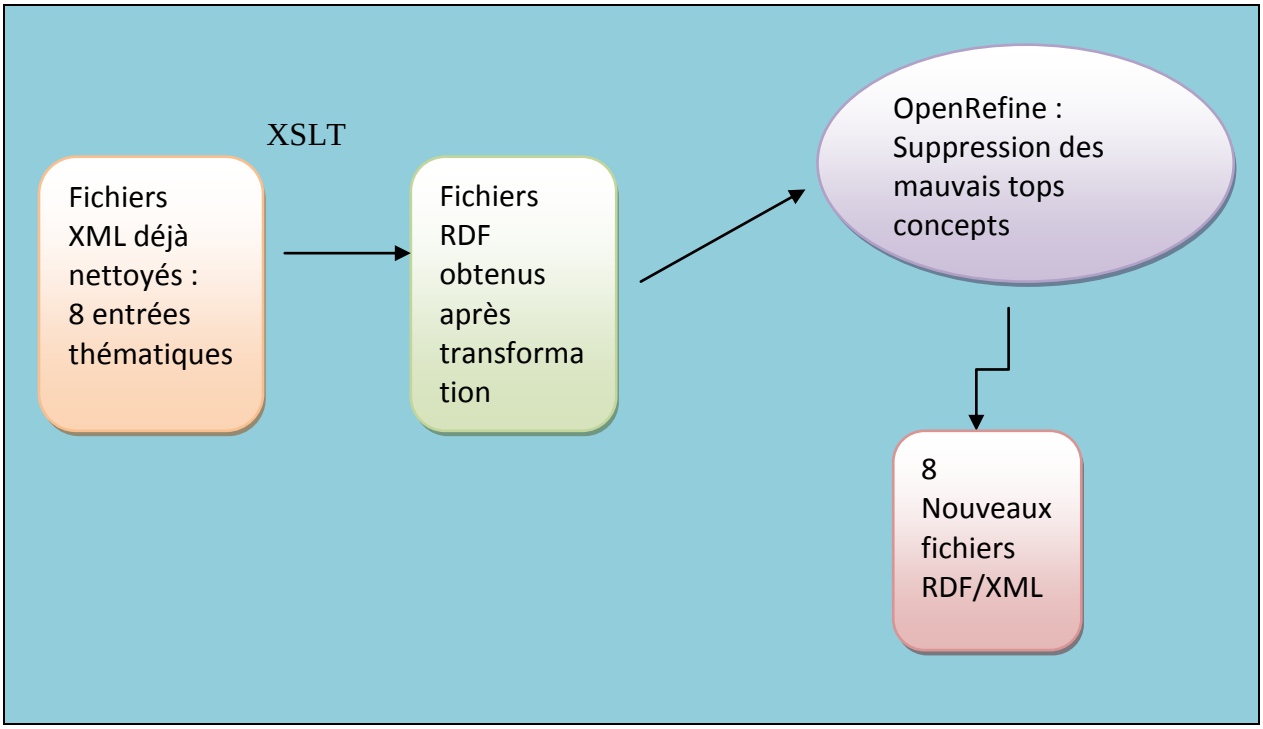

Figure 15 tableau récapitulatif des étapes de la transformation

```
@prefix Skosxl: <http://www.w3.org/2008/05/Skos-xl#> .
@prefix dc11: <http://purl.org/dc/elements/1.1/> .
@prefix ns0: <http://art.uniroma2.it/ontologies/VocBench#> .
@prefix Skos: <http://www.w3.org/2004/02/Skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dc: <http://purl.org/dc/terms/> .
<http://opendata.inra.fr/VocINRA/7266>
 a Skosxl:Label ;
 dc11:source "ProdINRA1" ;
 ns0:hasStatus "Validated" ;
 Skos:notation "7266"^^xsd:int ;
 dc:creator "110362J" ;
 Skosxl:literalForm "entre noeud"@fr .
<http://opendata.inra.fr/VocINRA/100000342>
 Skosxl:prefLabel <http://opendata.inra.fr/VocINRA/342>, <http://opendata.inra.fr/VocINRA/21260> ;
 Skos:prefLabel "adsorbant"@fr, "adsorbent"@en ;
 Skos:notation "100000342"^^xsd:int ;
 Skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 ns0:hasStatus "Validated" ;
 Skos:scopeNote "Objet d'étude" ;
 Skos:narrower <http://opendata.inra.fr/VocINRA/1000003997> ;
 a Skos:Concept ;
 Skos:topConceptOf <http://opendata.inra.fr/VocINRA/VocINRA> .
<http://opendata.inra.fr/VocINRA/1000003997> a Skos:Concept .
<http://opendata.inra.fr/VocINRA/21260>
 a Skosxl:Label ;
 ns0:hasStatus "Validated" ;
 Skos:notation "21260"^^xsd:int ;
 dc:creator "110362J" ;
 dc:created "2010-02-10T14:44:46.662+01:00" ;
 Skosxl:literalForm "adsorbent"@en .
<http://opendata.inra.fr/VocINRA/VocINRA> Skos:hasTopConcept <http://opendata.inra.fr/VocINRA/100000342> .
```

3.5 Évaluation et amélioration du résultat

Après avoir transformé nos fichiers XML en du SKOS, nous nous sommes intéressées à la qualité des données qui est d'une grande importance. La casse dans un thésaurus doit répondre aux normes qui le régissent dont nous parlerons dans le chapitre suivant. Dans notre, nouveau corpus, il y a les entités nommées qui constituent une entrée thématique numéro 5 (localisation géographique) mais sont en minuscule pour la plupart des termes. Il y a des accents qui sont omis lors de la création de certains termes par les documentalistes. Donc tout

ceci devrait être résolu avant le chargement dans VocBench. L'utilisation d'un outil permettant de traiter cette masse de données est nécessaire. OpenRefine et Notepad++ permettent de résoudre ces problèmes. Dans notre cas, nous avons utilisé l'onglet « text facet » dans OpenRefine pour mettre les majuscules aux termes de l'entrée thématique 5.

Pour les accents, nous avons suivi un peu suivi les approches adoptées par Zweigenbaum dans son article coécrit avec Natalia Grabar et intitulé *Accentuation des mots inconnus : application au thésaurus biomédical MeSH.* Ils se sont servis d'un programme en Perl mais nous avons repris ces approches avec le langage grell dans OpenRefine qui utilise les expressions régulières. Si un mot se termine par *eme,* le premier *e* du eme doit avoir un accent par exemple *problème*. Ce sera le cas pour deux *ee* comme *levee* l'avant dernier e doit être accentué. Nous avons aussi, les terminaisons en *ieme, iere, ere, atre* où il manque l'accent sur le premier *e* et sur le *a*.

Comme notre thésaurus est multilingue, l'accentuation des mots en se basant sur ces approches pose un peu problème et nécessite d'être vigilant. En effet, certains mots en anglais comportent des *ee* mais l'utilisation du filtre dans OpenRefine a permis de ne prendre en compte que les termes en français présents dans les colonnes « prefLabel » et « altLabel ». Il est nécessaire de préciser que nous avions des colonnes « prefLabel » et « altLabel » par langue. Néanmoins nous avons détecté quelques erreurs dans nos résultats car parmi les termes français choisis il y avait des termes anglais. En fait, il y a eu une erreur de choix de langue lors de l'indexation car après vérification, ces mots avaient pour langue le français dans le référentiel des mots-clés. Certains mots sont alors indexés avec un mauvais choix de la langue. C'est le cas avec les termes *bee, honeybee* qui sont des termes anglais. Ce problème illustre bien l'importance de la qualité des données d'origine.

Les signes de ponctuation sont souvent ajoutés à certains termes français comme les tirets pour la plupart des mots composés, l'apostrophe pour certains mots comme « *respect de l'environnement* ». Nous n'avons pas remarqué d'anomalies au niveau des mots avec des apostrophes quand nous avons fait un filtre, toujours dans OpenRefine, sur les mots composés avec *de* suivi de *l* nous n'avons rien trouvé. Pour les tirets par contre, nous en avons corrigé dans l'entrée thématique 5 (Localisation Géographique) qui comportent des noms de Pays. Mais nous avons vu qu'avec les normes AFNOR, le tiret n'était obligatoire que quand il avait un impact dans le sens d'un mot au cas où il serait omis.

Nous avons traité aussi le cas des minuscules pour les entités nommées composant en majorité l'entrée thématique 5. Comme tout nom propre commence par un une majuscule, nous avons mis la majuscule au début de chaque terme si c'est un terme simple. Pour les

entités nommés qui sont des mots composés, le début de chaque mot sauf les articles. Par exemple : *Afrique du Sud, Europe de l'Est.*

Par ailleurs, il est important de rappeler qu'un squelette a été développé pour pouvoir exporter nos données en RDF depuis Open Refine.

4 Les relations sémantiques

Le langage naturel est considéré comme ambigu et n'aiderait pas en tant que tel dans l'indexation des documents qui nécessite une certaine précision dans le choix des termes. Le problème de la synonymie, de polysémie et d'homonymie sont contournables avec l'élaboration d'un thésaurus. Dans ce dernier, les mots-clés doivent être répertoriés en établissant des liens entre eux. Les relations que peuvent entretenir ces termes sont multiples.

4.1 La relation hiérarchique

D'après Michel Hudon, « La relation hiérarchique facilite la navigation verticale dans le thésaurus, permettant à l'utilisateur d'atteindre le degré de précision désiré dans l'indexation ou la recherche. » [HUDON, p.57]

La relation hiérarchique qui est aussi appelée «relation d'inclusion» établit un lien entre les termes en partant de celui qui a un plus haut niveau pour en faire un terme préférentiel générique (**TG**) d'un autre terme préférentiel qui sera à son tour un spécifique (**TS**). Nous nous rendons compte qu'il crée des ensembles qui regroupent les génériques et des sous ensembles qui se composent de termes spécifiques. Ces derniers se retrouvent dans le champ sémantique des termes génériques qui représentent à leur tour les catégories générales, les entités à large sens.

Dans notre thésaurus écrit avec du SKOS, les termes qui entretiennent une relation hiérarchique sont marquée par « Skos: broader » pour les termes génériques et sa relation inverse « Skos: narrower » pour les termes spécifiques.

```
<http://opendata.inra.fr/VocINRA/10000093422>
 a Skos:Concept ;
 Skos:notation "10000093422" ;
 Skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 Skosxl:prefLabel <http://opendata.inra.fr/VocINRA/93422> ;
 Skos:prefLabel "aléa climatique"@fr ;
 Skos:scopeNote "Question sociétale" ;
 ns0:hasStatus "Validated" ;
 Skos:broader <http://opendata.inra.fr/VocINRA/100000101> .
```

```
 a Skos:Concept ;
 Skos:notation "100000101" ;
 Skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 Skos:scopeNote "Question sociétale" ;
 ns0:hasStatus "Validated" ;
 Skosxl:prefLabel  <http://opendata.inra.fr/VocINRA/101>,  <http://opendata.inra.fr/VocINRA/21118> ;
 Skos:prefLabel "accident climatique"@fr, "climatic hazards"@en ;
 Skosxl:altLabel <http://opendata.inra.fr/VocINRA/21119> ;
 Skos:altLabel "weather hazard"@en ;
 Skos:narrower <http://opendata.inra.fr/VocINRA/10000093422> ;
 Skos:topConceptOf <http://opendata.inra.fr/VocINRA/VocINRA> .
```

Dans ces exemples, le terme générique (TG) *aléa climatique* est et le terme spécifique (TS) est *accident climatique*.

Nous avons, dans notre thésaurus, des termes génériques et des termes spécifiques. Il est important de souligner que parmi ces termes spécifiques il y en a qui sont également eux-mêmes des termes génériques.

Dans la relation hiérarchique, nous avons une relation tête qui figure au plus haut niveau, seuls les termes qui n'ont pas de génériques peuvent être des têtes hiérarchiques dans un thésaurus. Ce sont les « tops concepts » (**TT**) notés dans notre thésaurus par « topConceptOf » et son inverse « hasTopConcept », relations définies entre un concept et le thésaurus (ou Skos :ConceptScheme).

```
<http://opendata.inra.fr/VocINRA/100000573>
 a Skos:Concept ;
 Skos:notation "100000573" ;
 Skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 Skos:scopeNote "Question sociétale" ;
 ns0:hasStatus "Validated" ;
 Skosxl:prefLabel <http://opendata.inra.fr/VocINRA/573>, <http://opendata.inra.fr/VocINRA/21365> ;
 Skos: prefLabel "aide sociale"@fr, "public assistance"@en ;
 Skosxl:altLabel <http://opendata.inra.fr/VocINRA/21366>, <http://opendata.inra.fr/VocINRA/21367> ;
 Skos:altLabel "social aid"@en, "social help"@en ;
 Skos:narrower <http://opendata.inra.fr/VocINRA/1000004966> ;
 Skos:topConceptOf <http://opendata.inra.fr/VocINRA/VocINRA> .
```

4.2 La relation associative

Définie par la norme ISO 25964 comme étant « la relation entre une paire de concepts qui ne sont pas reliés hiérarchiquement, mais partagent une forte connexion sémantique », la relation associative (**TA**) permet de créer des liens entre les concepts sémantiquement proches et souvent assimilés mentalement par les utilisateurs. Ces concepts associés n'appartiennent pas à la même hiérarchie mais ils doivent avoir des niveaux de spécificité comparables. L'association des concepts dépend du concepteur du thésaurus donc elle est subjective. Le concepteur doit veiller à l'utilité d'une telle association pour essayer de répondre aux besoins de l'utilisateur qui soumet une requête et lui permettre de décliner le champ sémantique du descripteur et sa définition.

M. Hudon dénombre 13 principaux types de relations associatives [HUDON , p. 132-133] que nous allons illustrer avec des cas semblables à partir de notre vocabulaire:

- la cause et l'effet : *Soleil et Énergie solaire*
- un tout et une composante essentielle : *Pollution atmosphérique et Ozone*
- une action et son agent : *Agriculture et agriculteur*
- une action et son produit : *Fertilisation et Azote*

- une action et son objet : *Financement public et dépense globale*
- une action et le lieu de son déroulement : *Essai en laboratoire et Laboratoire*
- une science et son objet : *Agronomie et Culture*
- un objet et sa propriété : *Poison et Toxicité*
- un objet et son application : *Ordinateur et Traitement informatisé de données*
- un objet et un de ses matériaux constitutifs : *Cuir et Animal*
- des concepts de sens proche : *Disette et Famine*
- des antonymes : *Sol cultivé et non cultivé*
- des concepts complémentaires : *Enseignement et Apprentissage*

```
<http://opendata.inra.fr/VocINRA/1000007893>
 a Skos:Concept ;
 Skos:notation "1000007893" ;
 Skos:inScheme <http://opendata.inra.fr/VocINRA/VocINRA> ;
 Skos:scopeNote "Question sociétale" ;
 ns0:hasStatus "Validated" ;
 Skosxl:prefLabel <http://opendata.inra.fr/VocINRA/7893>, <http://opendata.inra.fr/VocINRA/88730> ;
 Skos:prefLabel "famine"@fr, "food shortage"@en ;
 Skosxl:altLabel <http://opendata.inra.fr/VocINRA/25026>, <http://opendata.inra.fr/VocINRA/25027>, <http://opendata.inra.fr/VocINRA/88729> ;
 Skos:altLabel "disette"@fr, "faim (problème socioéconomique)"@fr, "pénurie alimentaire"@fr ;
 Skos:broader <http://opendata.inra.fr/VocINRA/1000003388> .
<http://opendata.inra.fr/VocINRA/7893>
 a Skosxl:Label ;
 Skosxl:literalForm "famine"@fr ;
 Skos:notation "7893" ;
 ns0:hasStatus "Validated" ;
 dc11:source "ProdINRA1", "Termsciences" ;
 dc:creator "110362J" .
```

Certains de ces exemples cités plus hauts ne sont pas reliés entre eux. Ils sont repris dans le thésaurus comme des termes préférentiels même s'ils sont sémantiquement liés. Or, si nous nous basons sur la nouvelle norme, ces termes devraient être reliés.

Dans l'IMC, les relations associatives sont représentées par la balise « associe » pour tous les types de relations associatives existantes. Il n'y a pas de distinction ni de précision exacte entre leur type d'association. Quand les données sont transformées, les termes associés sont

devenus des « altLabel » de même que les termes traduits. Nous nous sommes rendu compte que les relations ne sont pas bien définies au préalable, c'est-à-dire si deux termes qui devraient être associés autour d'un même concept ne le sont pas, l'ambigüité va toujours demeurer.

4.3 La relation d'équivalence

La norme ISO 25964 définit la relation d'équivalence comme étant « la relation entre deux termes d'un thésaurus représentant chacun le même concept » [Clause 2.18].

Elle permet d'établir un lien entre les termes qui sont proches c'est à dire qui peuvent exprimer la même idée ou renvoyer au même référent. Elle est souvent notée par **EM** ou **EP.**

Dans l'IMC, les relations d'équivalence sont définies par la notation « *Skos :altLabel* »

Il existe trois types de relation d'équivalence:

- **Synonymie véritable**

La forme complète d'un mot et son abréviation : *Institut National de Recherche Agronomique* et *INRA*. Seule l'abréviation est présente dans le vocabulaire, la forme complète n'y est pas associée.

Le nom vernaculaire et le nom scientifique sous lesquels un même concept est connu : *abricotier et Prunus armeniaca*. Dans notre vocabulaire, ces deux termes sont considérés tous les deux comme des « prefLabel». La relation « altLabel » n'est pas définie or il devrait l'être.

Les variantes orthographiques comme l'appellation ancienne et l'appellation moderne d'un même concept : Zaïre *et République Démocratique du Congo.*

Les termes d'origine linguistique différente lorsqu'ils sont couramment utilisés pour désigner un même concept dans une seule langue : *feedback et rétroaction.*

Les termes d'origine culturelle différente lorsqu'ils sont couramment utilisés pour désigner un même concept dans une seule langue : Bayern *et Bavière.*

- **Quasi-synonymie**

Ce sont des termes dont leur niveau hiérarchique diffère et sont spécifiques pour être employés comme des descripteurs. Mais ils sont aussi nécessaires à l'exhaustivité d'un thésaurus. Le terme qui est le plus général entre deux quasi-synonymes sera choisi comme descripteur.

Dans VocINRA, nous avons aussi des quasi-synonymes même s'ils ne sont pas reliés : *Adaptation et Accoutumance*

Il y a aussi les termes employés pour des concepts assez proches mais difficile ou impossible pour un non spécialiste de les repérer.

- **Antonymie**

Des termes antonymes sont souvent considérés comme équivalents quant ils représentent deux pôles d'un même axe sémantique comme *Imposition et Non imposition* qui ne sont pas reliés dans notre vocabulaire.

Notation Générale	**Notation Skos**
TG	**Skos :broader**
TA	**altLabel**
EP	**altLabel**
EM	**prefLabel**
TT	**topConceptOf**
TS	**Skos : narrower**

Figure 16 liste des termes employés pour définir les relations sémantiques

5 Extraction des synonymes

Cette partie est consacrée à la méthode adoptée pour regrouper les termes qui sont synonymes ou partageant le même concept pour mieux structurer notre terminologie. Beaucoup de termes ont des relations sémantiques non définies ce qui fait que la plupart des termes sont des Tops Concepts or ils ne devraient pas l'être. Même si nous envisageons de faire de nos entrées thématiques des tops concept de VocINRA plus tard, il faudra quand même que ces termes qui ne sont pas reliés le soient car un thésaurus doit lever l'ambiguïté du langage naturel comme la synonymie. Nous ne devons pas déroger à cette règle si nous voulons que notre thésaurus réponde aux normes de l'ISO 25964.

Nous nous sommes basés de quelques travaux de Thierry Hamon pour inférer des liens de synonymies entre les termes [Hamon et al, 1999]. Trois conditions sont posées et si l'une est présente dans sa terminologie, le lien de synonymie entre les termes peut être établi entre ces termes :

• Règle 1: les têtes sont identiques et les expansions sont synonymes;
• Règle 2: les têtes sont synonymes et les expansions sont identiques;
• Règle 3: les têtes sont synonymes et les expansions sont synonymes.

Pour obtenir un résultat qui permettrait d'appliquer ces règles, il faudra mettre en place un filtrage qui va permettre d'extraire tous les termes qui ont une tête identique ou une expansion identique.

Nous avons écrit à nouveau un programme en java qui va nous aider à retrouver tous les termes qui ont une tête ou une expansion identique. Le programme va extraire les termes sans tenir compte de la langue des termes. Il est important de préciser que ce type de filtrage ne s'applique qu'aux termes composés. Or, dans notre thésaurus nous avons aussi des termes simples parmi lesquels certains sont synonymes. Un programme est nécessaire pour traiter des termes simples afin de les regrouper. Malheureusement nous n'avons pas eu le temps de créer un programme qui allait nous permettre de faire un filtrage des termes simples à partir d'un dictionnaire technique avec des synonymes. Ce programme s'il avait été réalisé allait récupérer toutes les relations sémantiques entre les termes simples définies dans le dictionnaire choisi.

Néanmoins, avec le programme d'extraction des expansions et de têtes nous avons obtenu un nombre important de termes ayant les mêmes expansions ou têtes. Pour la plupart, ils apparaissent comme des tops concept dans notre vocabulaire or ils pouvaient être liés.

De notre part, nous suggérons que dans un premier temps, tous les termes ayant une tête identique soient regroupés autour d'un même concept avec la même tête dans le fichier XML si possible. Ainsi, la relation hiérarchique sera modifiée car certains termes vont passer de *top concept* à *narrower*.

Dans un second temps, il faudra revenir à la liste des termes extraits pour essayer de repérer les termes ayant une même tête à ne pas regrouper avec les autres de termes avec les mêmes critères.

Une fois ces termes repérés, les modifications pourront être apportées à partir de VocBench qui permet de déplacer des termes.

Par ailleurs, les termes avec une expansion identique, si relation d'équivalence entre les termes simples était définie au préalable, il serait plus simple de les regrouper car il y a des termes qui sont équivalents que nous ignorons. Seul un programme d'extraction expliqué précédemment pourra les détecter. Pour certains, ils sont faciles à relier car leur équivalence est connu même s'ils ne sont pas forcément établis dans le vocabulaire.

Par exemple les termes « *méthode* » et « *technique* » sont synonymes, deux mots ayant des expansions identiques avec pour têtes ces deux termes comme méthode de et technique de , il en sera déduit « *technique de dosage* » et « *méthode de dosage* » sont équivalents. L'un peut être déclaré comme *altLabel* de l'autre.

contrat travail
contrat de travail
division travail
division du travail
économie du travail
économie travail
organisation de travail
organisation du travail

agriculture à plein temps
agriculture à temps partiel
agriculture alternative
agriculture compétitive
agriculture contractuelle
agriculture de complement
agriculture de conservation
agriculture de subsistance
agriculture économe
agriculture énergétique

Figure 17 extrait résultat sur les termes avec les mêmes têtes

6 Alignement

Le partage des données pour qu'elles soient exploitables et interopérables passe par le lien de ses données à partir desquelles d'autres données pourront être régénérées. Souvent, ces données même ayant des sources différentes peuvent partager des connaissances semblables. Si nous prenons le cas de deux vocabulaires contrôlés conçus pour le même domaine forcément il va y avoir des concepts qui sont quasiment ou totalement les mêmes. Donc pour identifier ces concepts en commun et enregistrer leur relation, l'alignement sera préconisé.

L'alignement entre concepts est une méthode visant à créer une relation entre deux concepts de vocabulaires différents. La relation peut s'agir d'une équivalence totale ou approximative. L'équivalence totale est établie entre deux concepts qui renvoient aux mêmes notions et l'équivalence approximative porte deux termes. Tous les thésaurus peuvent s'aligner entre eux certes mais il n'est pas recommandé car s'aligner avec un thésaurus d'un domaine différent ne fournira pas de bons résultats.

Notre vocabulaire spécialisé dans l'agronomie et aux questions qui sont en rapport va être aligné aux vocabulaires du même domaine, à savoir, GACS et AGROVOC et qui sont un schéma de concept SKOS.

Skos offre des propriétés pour définir des alignements hiérarchiques avec *Skos :broadMatch, Skos :narrowMatch,* **associatif** avec Skos :relatedMatch. Si deux concepts sont similaires peuvent être interchangeables la propriété « close :Match est utilisée. L'exact :Match aligne deux concepts qui sont des synonymes exacts.

Notre alignement n'a pas totalement abouti car au moment de la rédaction de ce document nous étions à la phase test des fonctionnalités d'OpenRefine et VocBench et à la recherche d'autres outils qui vont nous permettre de s'aligner :

OpenRefine

Dans OpenRefine, il y a une fonctionnalité qui permet de faire des alignements, elle se nomme *reconcile* et peur être associé à l'extension RDF installée à l'avance dans le workspace d'OpenRefine. Elle est basée sur des APIs ou du SparqlEndpoints.[16] Nous avons testé l'alignement dans OpenRefine pour réconcilier les deux vocabulaires mais cela n'a pas marché. En effet, la réconciliation s'est faite sur la colonne des « prefLabel » en français. Il n'avait trouvé aucun concept semblable mais quand nous avons essayé la réconciliation avec le NCBI taxonomy il a trouvé des concepts communs. Nous pensons que le fichier extrait d'AGROVOC ne permet pas de faire l'alignement.

[16] http://wiki.inra.fr/wiki/traitementsdocumentaires/Main/OpenRefine

VocBench

VocBench permet aussi d'aligner des concepts grâce à la fonctionnalité *alignement*. L'alignement dans VocBench est manuel c'est-à-dire il se réalise par concept il ne peut pas être fait avec tous les concepts en même temps.

```
@prefix Skosxl: <http://www.w3.org/2008/05/Skos-xl#> .
@prefix Skos: <http://www.w3.org/2004/02/Skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

<http://id.agrisemantics.org/gacs/xl_lo_3a6e6aeb> a Skosxl:Label .
<http://id.agrisemantics.org/gacs/xl_de_966ad402> a Skosxl:Label .
<http://id.agrisemantics.org/gacs/C31771>
  Skos:prefLabel "プルーン"@ja, "pruimedanten"@nl, "sušené slivky"@sk, "sušené švestky"
@cs, "ciruela pasa"@es, "برقوق مجفف"@ar, "чернослив"@ru, "erik kurusu"@tr, "گوجه‌های برتاذ
ی"@fa, "Pruneau"@fr, "ลูกพรุน"@th, "ameixas secas"@pt, "Prugne secche"@it, "말린자두"@
ko, "prunes"@en, "李脯"@zh, "Śliwka suszona"@pl, "प्रून्स (सूखे आड़ू)"@hi, "Zwetsche"@de,
 "aszalt szilva"@hu ;
  a <http://id.agrisemantics.org/vocab#Product>, Skos:Concept ;
  Skos:broader <http://id.agrisemantics.org/gacs/C17481> .

<http://id.agrisemantics.org/gacs/xl_en_ead6a17f> a Skosxl:Label .
<http://id.agrisemantics.org/gacs/xl_hi_66d770c8> a Skosxl:Label .
<http://id.agrisemantics.org/gacs/xl_nl_eec18108> a Skosxl:Label .
<http://id.agrisemantics.org/gacs/C1939>
  a Skos:Concept, <http://id.agrisemantics.org/vocab#Product> ;
  Skos: prefLabel "果品"@zh, "Przetwory owocowe"@pl, "gyümölcstermék"@hu, "فراورده‌های
میوه "@fa, "Produto à base de fruta"@pt, "పండ్ల ఉత్పాదితాలు"@te, "ຜະລິດຕະພັນໝາກໄມ້"@l
o, "果実加工品"@ja, "과일생산물"@ko, "Prodotti a base di frutta"@it, "ผลิตภัณฑ์แปรรูปจากผลไม้"
@th, "fruit products"@en, "fruitproducten"@nl, "फल उत्पाद"@hi, "منتجات الفواكه"@ar, "meyve
(ürün)"@tr, "ovocné výrobky"@sk, "ovocné výrobky"@cs, "плодово-ягодная продукция"@
ru, "productos derivados de las frutas"@es, "Obst"@de, "produse din fructe"@mo, "Produit à
base de fruits"@fr ;
  Skos:narrower <http://id.agrisemantics.org/gacs/C17481> .
```

<http://id.agrisemantics.org/gacs/C17481>

Skos:prefLabel "Trockenobst"@de, "Fruits séchés"@fr, "乾燥果物"@ja, "Owoc suszony"@pl, "فواكه مجففة"@ar, "ผลไม้แห้ง"@th, "میوه‌های خشک‌شده"@fa, "aszalt gyümölcs"@hu, "сухофрукты"@ru, "건과"@ko, "果干"@zh, "gedroogd fruit"@nl, "frutas secas"@es, "frutas secas"@pt, "ໝາກໄມ້ແຫ້ງ"@lo, "sušené ovocie"@sk, "kurutulmuş meyve"@tr, "sušené ovoce"@cs, "शुष्क फल"@hi, "fructe uscate"@mo, "dried fruit"@en, "Frutti essicati"@it ;

Skos:exactMatch <http://lod.nal.usda.gov/nalt/34533>, <http://aims.fao.org/aos/AGROVOC/c_2382>, <http://id.cabi.org/cabt/40756> ;

Alignement du concept *fruits secs* dans GACS avec CABI et AGROVOC

Figure 18 fichier RDF extrait de GACS

URL de l'extrait : http://tester-os-kktest.lib.helsinki.fi/gacsdemo/gacs/fr/page/?uri=http%3A%2F%2Fid.agrisemantics.org%2Fgacs%2FC17481&

7 Conclusion et Perspectives

Nous avons travaillé sur le vocabulaire de l'INRA qui était au départ un ensemble de fichiers XML que nous avons transformé grâce à une feuille de style pour obtenir un vocabulaire en un schéma de concepts SKOS. Cette transformation était un passage d'un système non conforme au partage des données vers un système qui va rendre nos données interopérables. Le logiciel VocBench est capable de gérer notre thésaurus mais certains points sont à préciser. En effet, pour la réorganisation de l'arborescence, il peut être utile car il y a la fonctionnalité « move concept » pour déplacer les concepts mais c'est coûteux en temps car le déplacement ne peut pas se faire en bloc, c'est-à-dire il faut déplacer les concepts un par un. Donc, il est préférable de privilégier la commande sparql update qui aide à modifier des triplets RDF.

Nous avons pu revoir la qualité de nos données, ce qui nous a permis de déceler dans nos fichiers des anomalies d'ordre informatique au niveau des identifiants des entrées thématiques (identifiant0). Ce problème semait une confusion car nous ne pouvions plus distinguer dans VocBench les concepts avec des termes *deprecated* ou *proposed* car ils s'affichaient tous avec la même couleur sans suivre la légende définie dans VocBench.

Durant ce travail, nous nous sommes intéressés à la qualité des données en se basant sur la Norme 25964. Nous avons remarqué que beaucoup de termes étaient au pluriel. Pour certains des termes c'était normal car ils sont employés naturellement au pluriel en langage naturel mais pour d'autres le singulier est une « obligation » si nous voulons nous conformer à la norme précitée. Nous avons aussi réussi à résoudre le problème des majuscules de l'entrée thématique 5 qui regroupe les entités nommées car elles étaient écrites pour la plupart en minuscules. Avant d'arriver au résultat, qui nous a permis de vérifier les erreurs qu'il y avait dans nos fichiers, un filtrage avait été fait sur les termes issus de ProdINRA qui nous fait penser que le traitement automatique des langues et le web sémantique peuvent être vus comme des disciplines complémentaires. En effet, sans la méthode d'extraction appliquée nous n'allions pas pouvoir réduire la taille de notre vocabulaire. Nous n'allions non plus pas pouvoir établir la liste des termes libres qui avaient les mêmes têtes et les mêmes expansions qui sont à regrouper autour d'un même concept. Pour cela, nous pensons que la liste obtenue peut être soumise à un spécialiste qui validera les termes à regrouper autour d'un même concept car avoir la même tête ou la même expansion ne signifie pas partager le même réseau sémantique. Comme cette extraction était basée sur les termes composées, une méthode permettant de trouver les termes simples qui sont synonymes est aussi à prévoir pour résoudre le problème sur les relations sémantiques car beaucoup de termes sont définis comme

prefLabel alors qu'ils pouvaient être reliées à d'autres termes afin d'établir le lien de prefLabel et d'altLabel entre eux.

Il y a aussi le cas du pluriel et du singulier qu'il faudra revoir afin de choisir si les autres termes qui devraient être au singulier pourront garder la même forme pluriel ou non dans le thésaurus.

Un autre point important que nous avons découvert en traitant les accents est le choix de la langue lors de l'indexation car il y a des mots anglais qui sont mis dans les termes français. Nous jugeons nécessaire de les détecter pour corriger cette erreur.

Certes tous les objectifs n'ont pas été atteints à savoir la réorganisation de l'arborescence et aussi l'alignement avec GACS et AGROVOC, mais une bonne partie du travail a été réalisé, notamment le test de VocBench pour continuer les corrections et des instructions et procédures disponibles pour refaire les transformations. Personnellement ce stage m'a permis de mieux comprendre la gestion des connaissances dans un institut de recherche, l'importance d'avoir un vocabulaire contrôlé et aussi les étapes à suivre dans l'élaboration d'un thésaurus répondant aux règles de l'ISO et suivant les recommandations du web sémantique.

Bibliographie

AMARGER, Fabien, et al. s. d. « État de l'art : Extraction d'information à partir de thésaurus pour générer une ontologie ».

BILODEAU, Benoît. 2012. « De la place des normes dans le thésaurus RASUQAM ». *Documentation et Bibliothèques* 58 (3): 141‑52.

CHARLET, Jean, Philippe LAUBLET, et Chantal REYNAUD. 2003. « Web sémantique Rapport final ».

CHICHEREAU, Dominique, Odile Contat, Danièle Dégez, Alina Deniau, Michèle Lénart, Claudine Masse, et Dominique Ménillet. 2009. « Les normes de conception, gestion et maintenance de thésaurus ». *Documentaliste-Sciences de l'Information* 44 (1): 66-74.

Cyrot, Catherine, et Christian Preuss. 2009. « Réingénierie de thésaurus : une étude de cas ». *Documentaliste-Sciences de l'Information* 46 (3): 4-13.

CYROT, Catherine, et Christian PREUSS. 2009. « Réingénierie de thésaurus : une étude de cas ». *Documentaliste-Sciences de l'Information* 46 (3): 4-13.

DALBIN, Sylvie. 2009a. « Thésaurus et informatique documentaires ». *Documentaliste-Sciences de l'Information* 44 (1): 42-55.

DALBIN, Sylvie. 2009b. « Thésaurus et informatique documentaires ». *Documentaliste-Sciences de l'Information* 44 (1): 76-80.

DALBIN, Sylvie. 2013. *Livre blanc ISO 25964-1 - Thésaurus pour la recherche documentaire*.

DEXTRE CLARKE, Stella G., LEI ZENG, Marcia. «Standard Spotlight: From ISO 2788 to ISO 25964: The Evolution of Thesaurus Standards towards Interoperability and Data Modeling. » Information Standards Quarterly, 2012, Winter 24(1):20-26

DORIA, Orélie Desfriches, et Manuel ZACKLAD. 2010. « Améliorer la recherche d'information à l'aide de thésaurus « ad hoc » ». *Document numérique* 13 (2): 13-40.

DUPUCH, Marie, et al. 2012. « Structuration de terminologies », 431-444.

ERMINE, Jean Louis. 2008. *Management et ingénierie des connaissances. Modèles et méthodes*.

FERRET, Olivier. 2010. « Similarité sémantique et extraction de synonymes à partir de corpus ».

FIESCHI, Marius, Pascal Staccini, Omar Bouhaddou, et Christian Lovis. 2010. *Risques, Technologies de l'Information pour les Pratiques Médicales: Comptes rendus des treizièmes Journées francophones d'informatique médicale, Nice, 28-30 avril 2009*. Springer Science & Business Media.

Gandon, Fabien, Catherine Faron-Zucker, et Olivier Corby. 2012. *Le web sémantique - Comment lier les données et les schémas sur le web ?* Paris: Dunod.

HAMON, Thierry, et al. 1999. « Détection de liens de synonymie: complémentarité des ressources générales et spécialisées », 61‑69.

HATCHUEL, Armand, Pascal Le Masson, et Benoît Weil. 2002. « De la gestion des connaissances aux organisations orientées conception ». *Revue internationale des sciences sociales*, n° 171: 29‑42.

HUDON, Michèle. 2009. *Guide pratique pour l'élaboration d'un thesaurus documentaire*. 2e éd. Montréal: Les Editions ASTED (diffusion France : ADBS).

KELLER, Loraine. 2013. « Encadrer la réingénierie d'un thesaurus : méthode, enjeux et impacts pour l'équipe d'un service de veille et documentation en entreprise ». http://memsic.ccsd.cnrs.fr/mem_00945542/document.

KISTER, Laurence, Evelyne Jacquey, et Bertrand Gaiffe. 2011. « Du thesaurus à l'onto-terminologie : relations sémantiques vs relations ontologiques ». *Corela. Cognition, représentation, langage*, n° 9‑1(juin). doi:10.4000/corela.1962.

MOUREAU, Magdeleine. 1968. « Problèmes posés par la structure d'un thesaurus ». Text. janvier 1. http://bbf.enssib.fr/consulter/bbf-1968-05-0201-001.

MOUREAU, Magdeleine. 1973. « Principe et développement d'un thesaurus ».

PARENTI, Luigi, et Daniela TISCORNIA. 1981. « Problèmes linguistiques et informatiques liés à l'établissement d'un thésaurus juridique ». In *Congrès international informatique et sciences humaines*.

RABAULT, Hélène, et Hélène ZYSMAN. 2011. « Les schémas de concepts et le Web sémantique : la norme sur les thésaurus ISO 25964 et le Web sémantique ». In .

« Savoirs CDI: L'évolution des normes et les outils de gestion des vocabulaires contrôlés ». 2016.

STACCINI, Pascal, Ali Harmel, Stéfan Darmoni, et Riadh Gouider. 2012. *Systèmes d'information pour l'amélioration de la qualité en santé: Comptes rendus des quatorzièmes Journées francophones*. Springer Science & Business Media.

STUDER, Rudi, Andreas HOTO, Gerd STUMME, et Raphael VOLZ. 2003. « Semantic Web - State of the Art and Future Directions. » 17 (3): 5.

SYLVA, Lyne Da. 2004. « Relations sémantiques pour l'indexation automatique ». *Document numérique* 8 (3): 135‑55.

VERBORGH, Ruben. 2013. *Using OpenRefine Free Download*.

Zweigenbaum, Pierre et Natalia Grabar. 2002. « *Accentuation des mots inconnus : application au thésaurus biomédical MeSH »*, 1-10.

Tableau des Figures