Génération automatique d'ontologie sur la valorisation de déchets organiques à partir des publications scientifiques

Landy Rajaonarivo[1,2], Christiane Rakotomalala[3,4],
Mathieu Roche[1,5], Sarah Valentin[1,5]

[1] INRAE, UMR TETIS, F-34398 Montpellier, France
[2] TETIS, Université de Montpellier, AgroParisTech, CIRAD, INRAE, Montpellier, France
[3] Recyclage et Risque, CIRAD, Université de Montpellier, Montpellier, France
[4] UPR Recyclage et Risque, CIRAD, Saint-Denis, La Réunion, France
[5] CIRAD, UMR TETIS, F-34398 Montpellier, France

Résumé

Ces dernières années, plusieurs pays se sont intéressés à la transformation des déchets organiques en raison des avantages que cette technique offre dans divers domaines, notamment l'économie, l'agriculture et l'environnement. Les pays du Sud ne font pas exception, même si les travaux et les techniques utilisés dans ces pays sont encore largement méconnus et ne sont pas décrits dans le vocabulaire du domaine. Comment pouvons-nous tirer parti des différentes approches étudiées et mises en œuvre dans ce domaine dans ces pays ? Dans cet article, une approche visant à tirer parti des méthodes de transformation des déchets est proposée, en tenant compte des articles scientifiques publiés dans ce domaine. Cette approche intègre le traitement automatique du langage naturel et l'ingénierie ontologique afin d'extraire non seulement des lexiques spécifiques au domaine mais aussi des entités spatiales, générant ainsi automatiquement une ontologie composée de ces lexiques spécifiques. L'idée est d'exploiter cette ontologie afin de pouvoir effectuer des analyses sémantiques à l'avenir. Les résultats ont montré que notre approche permet de découvrir des lexiques très spécifiques qui ne sont pas présents dans le vocabulaire du domaine et d'aligner les lexiques avec les ontologies connues dans le domaine.

Mots-clés

valorisation des déchets organiques, lexiques thématiques, sémantique, information spatiale, traitement de langue naturelle

Abstract

In recent years, several countries have taken an interest in the transformation of organic waste due to the advantages this technique offers in various areas, including the economy, agriculture, and the environment. Countries in the Global South are no exception, even though the work and techniques used in these countries are still largely unknown and are not described in the vocabulary of the field. How can we leverage the different approaches studied and implemented in this field in these countries ? In this article, an approach aimed at leveraging waste transformation methods is proposed, taking into account scientific articles published in this field. This approach integrates automatic natural language processing and ontological engineering in order to extract not only domain-specific lexicons but also spatial entities, thereby automatically generating an ontology composed of these domain-specific lexicons. The idea is to exploit this ontology to perform semantic analyses in the future. The results showed that our approach allows us to discover highly specific lexicons that are not present in the domain vocabulary and to align the lexicons with known ontologies in the domain.

Keywords

organic waste valorization, thematic lexicons, semantics, spatial information, natural language processing

1 Introduction

Ces dernières années, plusieurs pays ont manifesté un vif intérêt pour la valorisation des déchets organiques, en particulier les pays développés grâce à leurs ressources financières, matérielles et à leur expertise qui facilitent la mise en place et le développement des nouvelles techniques. Des techniques sont également mises au point dans les pays du Sud, bien qu'elles soient encore peu nombreuses et moins variées que celles utilisées dans les pays dits développés. Cependant, leur existence est souvent moins connue en raison d'un manque de communication ou de promotion. [13] indique, selon leur analyse, que l'état de recherche ou les cas d'études dans les pays du Sud restent peu connus. Selon le rapport de l'ONU-Habitat en 2025 [1], le monde génère entre 2,1 et 2,3 milliards de tonnes de déchets municipaux solides chaque année, allant des textiles et emballages aux appareils électroniques, plastiques et aliments. Mais les systèmes de gestion des déchets ont du mal à suivre le rythme. Un rapport de l'ONU-Habitat focalisé sur les pays en développement mentionne que les deux tiers des déchets ménagers et commerciaux sont des déchets organiques provenant

1. https ://unhabitat.org/international-day-of-zero-waste-2025

des cuisines. La valorisation de ces déchets organiques offre plusieurs avantages dans divers domaines notamment l'économie, la société, la santé publique, l'environnement, à savoir la réduction de pollution et des risques de propagation d'épidémies, la création d'emploi, la production d'énergie, l'amélioration du système agricole. De nombreux pays en développement ont manqué ces opportunités offertes par cette approche. Un moyen permettant de promouvoir l'approche de valorisation de déchets consiste à partager les techniques développées dans ces pays afin que chaque partie prenante (population locale, autorité locale, chercheur, entrepreneur, bailleur, etc.) puisse voir ce qui est faisable à son niveau (de petite à grande échelle) et de l'appliquer dans son domaine d'activité. Les parties prenantes peuvent collaborer ou s'inspirer des approches existantes selon les ressources dont elles disposent. C'est dans cette optique que ce travail de recherche a été mené. Dans ce contexte, l'objectif de nos travaux est de réaliser une analyse la plus exhaustive possible de la situation à travers les informations sémantiques contenues dans les articles scientifiques. Ainsi, dans un premier temps, des articles scientifiques ont été collectés sur la valorisation des déchets organiques dans les pays du Sud afin d'en extraire des lexiques spécifiques pour le domaine, de les normaliser et contextualiser en les représentant via une ontologie et enfin de les partager. L'objectif dans un second temps est de faire des analyses sémantiques sur la valorisation via cette ontologie. Ce papier décrit principalement les approches permettant d'atteindre ce premier objectif afin d'avoir des éléments clés pour l'analyse sémantique. Les questions que les parties prenantes peuvent se poser pour l'analyse sont les suivantes : Comment les techniques développées dans ce domaine ont-elles été réparties dans l'espace ? Existe-t-il des caractéristiques communes en termes de climat, d'économie, de société, de culture, etc., entre les lieux où les mêmes techniques ont été développées, en vue de transférer ces techniques vers d'autres lieux ? Les matériaux et les méthodes utilisés pour mettre en œuvre une technique diffèrent-ils d'une région à l'autre ? Les approches mises en œuvre ont-elles tendance à se concentrer davantage sur la théorie (discussion) ou sur la pratique (projets) ? La spécificité de notre approche réside dans les points suivants :

- Extraction des termes techniques du domaine et leur contextualisation non seulement au niveau thématique, mais aussi au niveau spatial
- Définition de l'intensité des liens (fort, faible, ou pas de lien) entre un lexique et les trois sujets tels que OWT (*Organic Waste Type*), TM (*Treatment Method*) et AV (*Agricultural Valorisation*)
- Génération automatique d'ontologie et son alignement vers des ontologies largement utilisées dans le domaine afin de respecter les principes FAIR (*Findable, Accessible, Interoperable, Reusable*).

La principale contribution de ce papier réside dans la valorisation et le partage de connaissances relatives au traitement des déchets organiques plutôt que dans ses aspects méthodologiques.

Le reste de ce papier s'organise comme suit : Section 2 présente l'État de l'art, Section 3 illustre le matériel et la méthode mise en œuvre, et Section 4 présente les résultats et discussions.

2 État de l'art

L'étude de la sémantique de données est déjà utilisée dans le domaine de la valorisation de déchets et bon nombre de ces études démontrent l'intérêt de la recherche sémantique dans ce domaine. [6] propose une approche d'analyse de relations sémantiques entre les articles dans la base de données Scopus sur le thème de la digestion agricole afin d'identifier les liens sémantiques existants entre les articles tels que la récurrence des mots-clés ou des thèmes de recherche. Le résultat de leur analyse a permis de montrer un intérêt croissant pour la technique de digestion agricole et le concept de bioraffinerie ces dernières années. L'approche décrite dans [3] vise à fournir des connaissances sur la valorisation des déchets issus de la bioraffinerie à des personnes qui ne sont pas expertes dans le domaine, en utilisant des approches d'ingénierie ontologique. [14] propose une approche de modélisation d'ontologies sur la valorisation de la bagasse de canne à sucre. Leur approche démontre la capacité à faire des raisonnements à l'aide de l'ontologie. Elle montre également la nécessité de l'utilisation d'ontologie pour le partage de données et la représentation de connaissance dans le domaine. [19] s'intéressent à la mise en connexion et à l'exploitation des données hétérogènes sur l'agriculture, l'environnement, l'alimentation et la santé qui sont liées à l'ingénierie alimentaire et aux bioproduits. Les auteurs indiquent que dans ce contexte, l'ontologie joue un rôle clé car elle fournit des représentations des connaissances et des structures formelles pour l'intégration de données. Une ontologie nommée PO2 (*Process and Observation Ontology*) a été proposée pour fournir des vocabulaires et des descriptions pour n'importe quelle procédure de traitement de biomasse et caractérise tous les composants en entrée et en sortie impliqués dans la procédure. Une grande partie des vocabulaires est venue de *European food classification system* (FoodEx2), *European Waste Catalog* (EWC) et d'autres nomenclatures internationales. La spécificité de leur approche réside dans la description structurée des ingrédients (intrants), des produits (extrants), des processus et des étapes, qui aide les parties prenantes (ex. les praticiens) à décrire leurs systèmes de production à l'aide des lexiques disponibles. Nous nous référons à cette ontologie pour créer la nôtre en considérant les composants et les méthodes ou les processus. Il a été indiqué dans [19] que, d'après les études de cas examinées, cette méthode innovante de sémantisation permet de répondre à des questions spécifiques posées par des experts dans ce domaine, à savoir prédire les performances du processus de microfiltration du lait dans un large éventail de conditions d'exploitation et de technologies membranaires. Une approche permettant la modélisation des ressources en déchets, en eau et en énergie via une ontologie, ainsi que des informations sur leur composition, leurs caractéristiques (chimiques et physiques) et les connaissances tacites relatives à leur flux, a été présentée

dans [18]. Il a été souligné dans [1] que la valorisation des déchets est l'un des thèmes de recherche importants de ces dernières années, mais que le vocabulaire utilisé dans ce domaine et ses définitions varient considérablement d'une ville ou d'un pays à l'autre. D'où l'intérêt de la mise en place d'un système de normalisation tel que l'ontologie. Bon nombre de ces approches se concentrent sur une technique spécifique dans le domaine de la valorisation des déchets, tandis que d'autres sont plus génériques et prennent en compte des techniques et des concepts beaucoup plus larges liés à l'ingénierie des déchets. Notre objectif est de créer une ontologie générique axée sur le domaine de la valorisation des déchets organiques, couvrant les composants (notés OWT : *Organic Waste Type*), les techniques de traitement (notées TM : *Treatment Method*) et les méthodes de valorisation (notées AV : *Agriculture Valorisation*) pris en compte dans ce domaine, et de définir l'intensité de lien entre chaque concept et ces trois composants s'il existe.

3 Matériel et Méthode

L'approche proposée est constituée de 5 modules principaux tels que : (i) la constitution du corpus, (ii) l'extraction, (iii) le filtrage et la structuration des données, (iv) l'enrichissement, ainsi que (v) la construction et la mise à jour de l'ontologie. Figure 1 illustre la procédure globale de notre approche de génération d'ontologie sur la transformation de déchets organiques en utilisant des données d'articles scientifiques. Cette approche se base sur trois techniques fondamentales à savoir le TALN (traitement automatique de langue naturelle), l'enrichissement des données (*data enrichment*) et la génération automatique d'ontologie (*ontology engineering*).

3.1 Constitution du corpus

À partir de plusieurs bases de données en ligne (WoS, Ovid, Scopus, Google Scholar, HAL, Cairn.info, AGRIS et Agritrop), nous avons constitué un corpus initial de 24 186 articles scientifiques publiés jusqu'en octobre 2021 et relatifs à la bio-transformation et à la valorisation des résidus organiques en agriculture dans les pays émergents et en développement. Les termes utilisés pour la recherche bibliographique dans chaque base de données, au moyen de requêtes spécifiques, sont détaillés dans les travaux de [15]. Le corpus a ensuite été réduit à 7692 documents après un premier tri des articles en anglais et un deuxième tri d'articles hors sujets à partir des titres, des mots-clés et des résumés. C'est sur ce corpus de 7692 références bibliographiques que l'extraction terminologique a été effectuée.

3.2 Extraction de données

3.2.1 Extraction de la terminologie et des informations thématiques associées

BioTex [12] a été utilisé pour réaliser une extraction automatique de 19 580 candidats-termes en anglais à partir des titres des articles de ce corpus. Cet outil, initialement développé pour l'extraction de termes biomédicaux, a été adapté à l'extraction de termes liés à la sécurité alimentaire [17]. BioTex effectue un filtrage linguistique par analyse syntaxique (nom-nom, adjectif-nom, etc.) puis intègre des mesures statistiques afin de classer les termes extraits. Notons qu'une étude en cours compare cette approche de fouille de textes avec des méthodes fondées sur les LLM (*Large Language Model*) pour l'extraction de la terminologie à partir de telles données textuelles.

3.2.2 Extraction spatiale

Les entités spatiales ont été extraites automatiquement du contenu textuel (titre et résumé), à l'aide du modèle GliNER *small* [20]. GliNER est un modèle pour l'extraction d'entités *zero-shot*, basé sur un encodeur transformateur bidirectionnel (de type BERT). Compte tenu du comportement de GliNER, qui extrait à la fois des entités spatiales absolues (toponymes précis tels que villes, régions ou pays) et des expressions géographiques plus générales ou descriptives (par exemple *subtropical countries*), un pré-traitement des entités spatiales candidates a été nécessaire afin de réduire le nombre d'expressions susceptibles de générer des erreurs de géocodage. Nous avons retiré : (1) les expressions ne commençant pas par une majuscule, (2) les expressions toutes en majuscules et de moins de 3 caractères, et (3) les expressions appartenant à une liste de *stop-words* manuellement construite sur la base des premiers résultats de l'extraction (ex. *Soil*, *Greenhouse*, etc.). Les entités candidates ont ensuite été normalisées en retirant les termes relatifs (*northwestern China* normalisé en *China*).

3.3 Filtrage et structuration des candidats-termes extraits

Les candidat-termes ou les lexiques thématiques ont été classés par thème en (1) types de résidus organiques (OWT :*Organic Waste Type*), (2) méthode de transformation (TM : *Treatment Method*) et (3) valorisation en agriculture (AV : *Agricultural Valorisation*). La pertinence a été établie suivant deux cycles itératifs d'évaluation par 6 experts du domaine sur un échantillon de 200 candidats-termes. Ils ont été classés en (i) très pertinents (+) quand ils sont directement associés à une ou plusieurs catégories, (ii) pertinents quand ils sont indirectement associés à une ou plusieurs catégories et (iii) non pertinents quand les termes ne sont associés à aucune des catégories. Cette procédure de pertinence a été ensuite appliquée par un expert à tout le reste des candidats-termes. Sur les 19580 termes candidats initiaux, environ 75% ont été manuellement annotés comme non pertinents qui sont sans lien avec les résidus organiques, la biotransformation ou la valorisation en agriculture. Parmi les 25 % de termes pertinents (4895), 2 079 étaient étroitement liés à la valorisation des résidus organiques dans les pays émergents et en développement, tels que les boues d'épuration, les eaux usées, le bétail, le fumier, les lisier, la digestion anaérobie, le compostage et le lombricompostage. Les 2079 termes pertinents sont relatifs à la valorisation des résidus organiques dans les pays émergents et en développement. Le protocole d'annotation est détaillé dans [15]. Plusieurs de ces termes figurent dans le

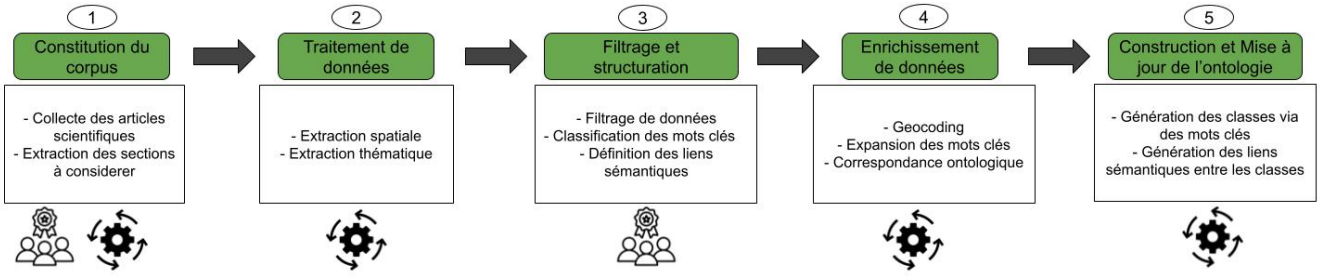


FIGURE 1 – Approche de génération d'ontologie via des mots-clés thématiques extraits des articles scientifiques sur la transformation de résidus

glossaire "*livestock and manure management 2011*"[2]. De plus, certains termes pertinents sont cités dans la littérature comme étant liés à la biotransformation et à la valorisation en agriculture. L'ensemble de ces termes est en libre accès sur un entrepôt de données Dataverse [16].

3.4 Enrichissement des données

L'enrichissement de données dans le cadre de ce travail consiste à enrichir les informations spatiales en utilisant la technique de Geo-coding afin d'avoir des informations détaillées et précises sur un lieu donné telles que son type (commune, département, pays), son adresse, etc. Par ailleurs, cet enrichissement a également consisté à étendre les mots-clés thématiques afin d'avoir les différents variants syntaxiques, à enrichir les mots-clés thématiques afin de connaître leur définition et leurs liens sémantiques avec d'autres mots-clés en utilisant la technique de correspondance ontologique. Ces techniques permettent non seulement de réduire le risque d'ambiguïté, mais également de favoriser l'interopérabilité et l'intégration des données avec des données ouvertes existantes dans le domaine (*Linked Open Data* comme Wikidata, AGROVOC et PO2, des ontologies ou thésaurus, etc.).

3.4.1 Géocodage

Les entités candidates et normalisées ont été géocodées grâce au gazettier Geonames [8], via la librairie *Geopy*. Nous avons conservé le premier résultat retourné, correspondant à l'entité ayant le score de pertinence le plus élevé selon l'algorithme de classement de GeoNames. Nous avons constaté que les expressions ne faisant pas référence à des noms de lieux explicites, mais à des formes adjectivales (ex : 'Brazilian'), pouvait générer des erreurs lors du géocodage automatique. Nous avons donc effectué un mapping automatique avec le pays correspondant. L'ensemble du processus d'extraction et de géocodage est disponible sur un dépôt Git[3].

3.4.2 Regroupement et expansion morpho-syntaxique

Nous avons appliqué une méthode d'expansion automatique des termes pertinents, consistant à associer, pour un terme donné (appelé "terme graine"), un ensemble d'expressions synonymiques correspondant à des variations morpho-syntactiques simples (allant de la plus simple, la forme plurielle, à des modifications de l'ordre des mots et/ou de l'insertion de conjonctions de type coordination). Par exemple, les expressions "*intensification of agricultural*" et "*intensification of agriculture*" sont des variantes morpho-syntactiques du terme graine "*agricultural intensification*". Cette approche d'identification des variantes présente un double avantage : elle permet d'identifier les diverses formes d'un terme dans les articles scientifiques, améliorant leur détection, et elle augmente les chances de trouver des correspondances lors de l'alignement ontologique à des fins de normalisation.

Nous avons tout d'abord constaté que la liste des 2079 termes candidats pertinents identifiés à l'issue de l'extraction terminologique contenait plusieurs fois le même terme graine, sous des formes morpho-syntactiques différentes. L'approche d'expansion terminologique a donc consisté en 2 étapes successives : (1) le regroupement des variantes synonymiques, sur la base de leur proximité morpho-syntaxique, dans la liste des 2079 termes pertinents identifiés à l'issue de l'extraction terminologique ; et (2) l'enrichissement de ces termes par l'ajout de nouvelles variantes morpho-syntaxiques. Nous avons ensuite enrichi cette représentation en recherchant, pour chaque terme graine, de nouvelles variantes morpho-syntactiques dans le corpus. Pour ces deux étapes successives, nous avons utilisé Fastr [9], un outil linguistique fondé sur des règles qui génère des variantes morpho-syntaxiques de termes à partir d'une liste de termes ou d'un corpus fourni en entrée. Pour le regroupement synonymique, nous avons appliqué Fastr sur tous les termes pertinents, puis validé manuellement les regroupements de termes graines/variantes proposés par l'outil. Pour l'expansion automatique, nous avons utilisé le corpus de 24 186 articles et généré automatiquement les variantes. Compte tenu du nombre conséquent de variantes obtenues, cette étape n'a pas été validée manuellement.

Les détails d'implémentation sont documentés dans un dépôt GitHub[4]. Notons enfin que des extensions sont en cours pour comparer et combiner une telle approche qui s'appuie sur des méthodes linguistiques avec des méthodes fondées sur les grands modèles de langues.

2. https ://ramiran.uvlf.sk/doc11/RAMIRAN%20Glossary_2011.pdf
3. https ://github.com/SarahVal/SciGeocoding
4. https ://github.com/SarahVal/FastrOrganicWastes

3.4.3 Correspondance ontologique

La correspondance ontologique a pour objectif de déterminer la correspondance des concepts ou des relations entre des ontologies. Plusieurs techniques ont été développées dans ce domaine [11], qui peuvent être classées en différentes catégories de méthodes, telles que celles fondées sur les noms, sur la similarité sémantique, sur la structure, etc. Ces méthodes peuvent aller de simples à sophistiquées. Dans le cadre de ce travail, nous adoptons l'utilisation de la méthode basique de correspondance de concepts basée sur le nom à partir de la liste de mots thématiques générée lors de l'extraction de la terminologie (voir section 3.2.1). La correspondance s'appuie sur la similitude exacte entre un concept d'une ontologie donnée et le label ou l'une des variantes du mot-clé thématique en question. Deux ontologies relatives à notre domaine d'études ont été considérées telles que le thésaurus multi-langues AGROVOC et l'ontologie PO2. AGROVOC [5] qui est développé par la FAO [6] est un ensemble de données ouvertes dédié à l'agriculture disponible pour un usage public [5] tandis que PO2 est une ontologie sur l'ingénierie alimentaire, fourragère, des bioproduits et des biodéchets pour l'intégration des données dans une bioéconomie circulaire et une approche axée sur les liens [19]. Notre choix repose sur le fait qu'AGROVOC définit un grand nombre de concepts sur l'agriculture et PO2 définit des concepts spécifiques liés à notre sujet d'étude, à savoir l'ingénierie de traitement des déchets organiques. La correspondance avec l'ontologie AGROVOC est réalisée de façon automatique à l'aide des requêtes SPARQL via une API d'AGROVOC [7] tandis que celle avec PO2 est réalisée à l'aide des requêtes SPARQL sur l'ontologie locale de PO2. À ce stade, nous disposons des résultats de mise en correspondance des mots-clés thématiques avec les concepts des deux ontologies considérées (voir Tableau 4). Il peut y avoir quatre scénarios possibles : un mot-clé thématique peut avoir des correspondances uniquement avec AGROVOC, uniquement avec PO2 (Figure 5), avec les deux (Figure 4), ou aucune (Figure 6).

3.5 Construction de l'ontologie

La construction et la mise à jour de l'ontologie ont été réalisées de manière entièrement automatique. Cela consiste en la génération des classes et des relations sémantiques. L'approche a été réalisée via des programmes en Python utilisant la bibliothèque *owlready2*. Dans le cadre de ce travail, nous avons pris en compte 204 mots-clés (ou lexiques thématiques) que nous considérons comme très pertinents afin de travailler dans un premier temps sur un petit ensemble de données.

3.5.1 Génération des classes

Un mot-clé thématique est représenté par une classe dans l'ontologie. Pour la structuration de l'ontologie, des classes parentes ont été d'abord créées telles que la classe «*agriculture field*» qui regroupera toutes les classes générées via les mots-clés thématiques, la classe «*organic waste type*» notée *OWT*, la classe «*treatment method*» notée *TM* et la classe «*agricultural valorisation*» notée *AV*. Une classe est caractérisée par les attributs suivants : id (l'identité), label, altLabel (les variantes), *seeAlso* (urls vers les ontologies externes qui définissent le mot-clé). L'attribut *seeAlso* utilise les résultats de la correspondance ontologique (voir section 3.4.3).

3.5.2 Génération des relations

Les relations se définissent comme des liens sémantiques entre deux classes dans l'ontologie. Nous avons considéré trois types de relations telles que la relation hiérarchique représentée par l'attribut *subClassOf* et les relations contextuelles, plus précisément, de relation de pertinence de liens qui sont représentées par les attributs *has_relation_with* et *has_strong_relation_with* et sont définis dans la section 3.3. Ces deux dernières relations sont de type *object properties* dont le domaine est la classe *agriculture field* et les *ranges* sont les classes OWT, TM et AV. Figure 4 illustre un exemple de ce type de lien : le concept *batch reactor* a un lien fort avec le concept *transformation method* selon le contexte. En d'autres termes, lorsque nous parlons d'un réacteur discontinu, nous faisons explicitement référence à une méthode de transformation.

4 Résultats et Discussion

Les résultats présentés dans cette section concernent l'enrichissement ou l'expansion des lexiques, l'extraction d'entités spatiales, ainsi que la génération et l'alignement d'ontologies.

4.1 Regroupement et expansion morpho-syntaxique

La première étape repose sur le regroupement des variantes morpho-syntaxiques extraites de la liste initiale de termes pertinents. Ce processus a conduit à l'identification d'une liste réduite de termes graines (nombres en gras dans la Table 1), chacun est associé à ses variantes correspondantes (nombres en italique dans la Table 1). Pour chaque regroupement, le terme-graine a été effectué manuellement par une experte du domaine.

La deuxième étape, consistant en la recherche de variantes sur le corpus d'articles, a permis d'extraire un total de 1907 à 3471 variantes par catégorie thématique.

Des exemples de résultats sont illustrés dans les Tables 2 et 3, l'ensemble des résultats du regroupement et de l'expansion automatique est disponible en libre accès sur un dépôt Davaverse [8].

4.2 Localisation

Sur les 2492 valeurs uniques d'entités spatiales candidates, 70,3% (n=1751/2462) ont été géocodées à l'issue de la chaîne de traitement, dont 198 à partir d'une forme adjectivale, 446 après nettoyage et 1059 directement à partir de leur valeur brute (Figure 2).

5. https ://www.fao.org/agrovoc/fr
6. Food and Agriculture Organization of the United Nations
7. http ://agrovoc.fao.org/sparql
8. https ://doi.org/10.18167/DVN1/G2IFIA

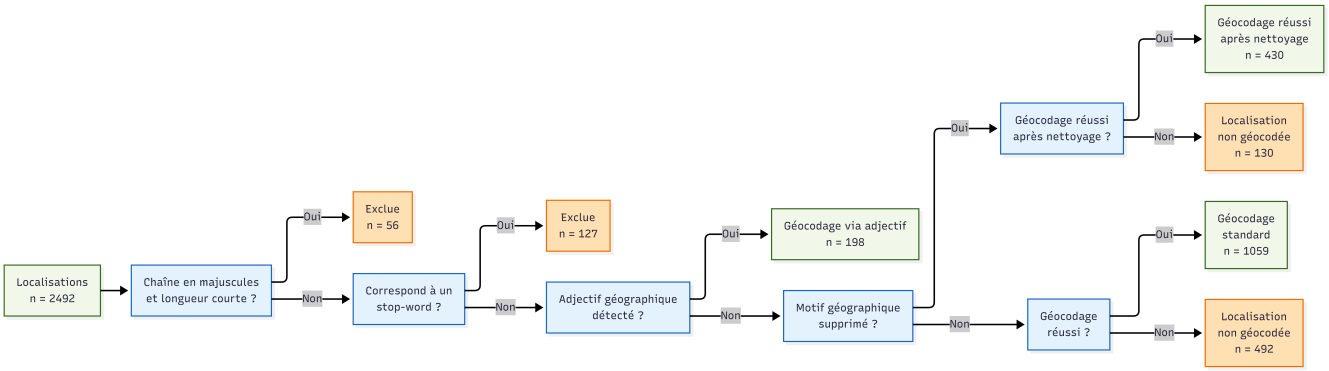
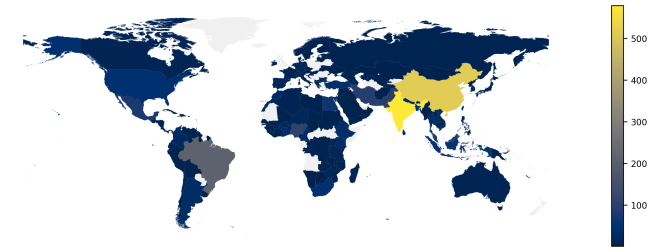
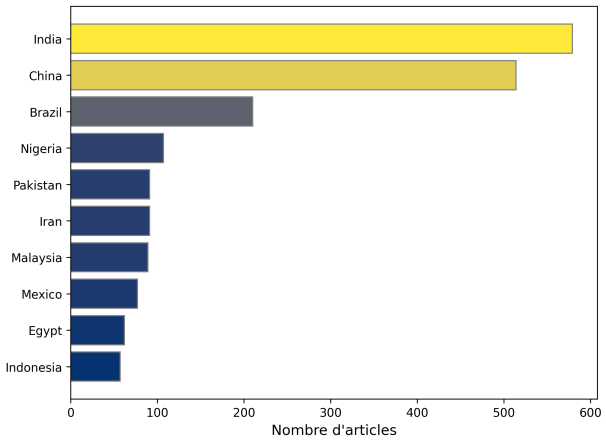


FIGURE 2 – Étapes de filtrage avant le géocodage des entités candidates



(a) Fréquence des articles par pays



(b) Top 10 des pays les plus fréquents

FIGURE 3 – Distribution du nombre d'articles par pays, obtenue par extraction automatique des entités géographiques et géocodage à partir du contenu textuel des articles.

TABLE 1 – Évolution du nombre de termes graines (en gras) et de variantes (en italique) après le regroupement et l'expansion morpho-syntactique à l'aide de Fastr et après validation manuelle.

Catégorie	Liste initiale	Regroupement	Expansion
TM+	**475**	**395**	**395**
	–	*80*	*1 907*
OWT+	**930**	**739**	**739**
	–	*191*	*3 321*
AV+	**674**	**581**	**581**
	–	*93*	*3 471*

TABLE 2 – Exemple de regroupements morpho-syntactiques (issus des termes pertinents)

Termes graines	**Variantes**
bio-waste	biowastes, bio-wastes, bio-waste
herb residue	herb residues, herbal residue, herbal residues
ligno-cellulose waste	lignocellulose waste, lignocellulosic wastes, lignocellulosic waste

Sur les 7692 articles, le processus d'extraction et de géocodage a permis d'attribuer un ou plusieurs pays à 2594 articles, soit 33,7% des articles. Les pays les plus représentés dans le corpus, sur la base des entités spatiales issues de leur contenu (titre et résumé), sont l'Inde, la Chine et le Brésil (Figure 3).

Nous avons réalisé deux évaluations afin d'analyser les résultats du géocodage. D'une part, l'ensemble des 2594 articles géocodés a été annoté manuellement, chaque article étant associé à un ou plusieurs pays correspondant à la ou aux zones d'étude abordées. Nous avons comparé les pays identifiés manuellement avec ceux obtenus à l'issue du géocodage. Pour 77,2% des articles (n = 2003/2594), les résultats obtenus automatiquement sont strictement identiques aux annotations manuelles. Par ailleurs, pour 92,1% des articles (n = 2390/2594), au moins un des pays identifiés au-

TABLE 3 – Exemple d'expansions morpho-syntactiques (issues du corpus)

Termes graines	**Variantes**
alley-cropping	alley cropped, alley cropping, cropped alley, alley crop, crop under alley, cropping systems alley
film mulch	film mulching, mulching film, mulch film, film mulched, mulch plastic film

tomatiquement correspond à un pays annoté manuellement. Nous avons ensuite analysé deux sous-échantillons :

- un ensemble de 50 pays "faux-positifs" (extraits automatiquement mais ne correspondant pas à l'annotation manuelle). Parmi eux, 48% (n=24/50) proviennent de l'extraction erronée d'entités spatiales (par exemple une confusion avec un nom de plante - *Euphorbia*, *Alium* - ou du vocabulaire technique - *reactor D*) qui retournent un résultat à l'issue du géocodage. Ensuite, 38% (n=19/50) proviennent d'une erreur de géocodage (notamment dans le cas d'entités spatiales couvrant plusieurs pays, telles que *lake Victoria*, *Mediterranean area*, ou des entités paysagères, telles que *Northern maize region*). Enfin, 14% (n=7/50) correspondent à des entités spatiales présentes dans le résumé, mais non reliées au pays d'étude, comme des mentions d'éditeur.
- un échantillon de 50 articles issus des articles non géocodés afin d'identifier d'éventuels faux négatifs, liés à la non-extraction d'une ou plusieurs entités spatiales. À l'issue de cette évaluation, aucune information spatiale n'a été relevée dans les titres ni dans les résumés des articles composant cet échantillon.

L'alignement des entités extraites avec GeoNames est une étape importante, permettant de relier des expressions textuelles potentiellement ambiguës à des entités géographiques identifiées, normalisées et intégrées dans une hiérarchie administrative explicite. Nos résultats montrent que le géocodage automatique s'avère globalement performant, avec une forte correspondance avec les annotations manuelles, tout en mettant en évidence certains verrous liés à l'extraction et à la géolocalisation d'entités spatiales à partir de résumés d'articles scientifiques. Tout d'abord, la majorité des articles (66,3%) sont non géo-localisables à partir du titre et du résumé seuls, en grande partie en raison de l'absence de mentions spatiales explicites. L'exploitation du texte intégral ou, à défaut, des informations d'affiliation des auteurs pourrait améliorer les performances de la géolocalisation. Cependant, cette approche dépend de l'accès libre aux articles, l'analyse du texte complet peut également introduire du bruit. D'autre part, nous avons constaté que les erreurs lors de l'extraction d'entités spatiales (entités non spatiales, ou entités spatiales paysagères, d'infrastructure, etc.) produisaient des erreurs nécessitant des étapes de filtrage. Dans ce contexte, deux pistes d'amélioration peuvent être envisagées : (1) la comparaison du modèle GliNER small à une version plus large ou plus récente, ainsi qu'à d'autres modèles de l'état de l'art (BERT [7], spaCy [2]) afin d'identifier la meilleure approche pour l'extraction et réduire le nombre d'entités extraites incorrectes et (2) intégrer une étape de désambiguisation ou de raisonnement spatial lors du géocodage [4], par exemple en explorant les performances d'approches de *prompting* avec de grands modèles de langage [21].

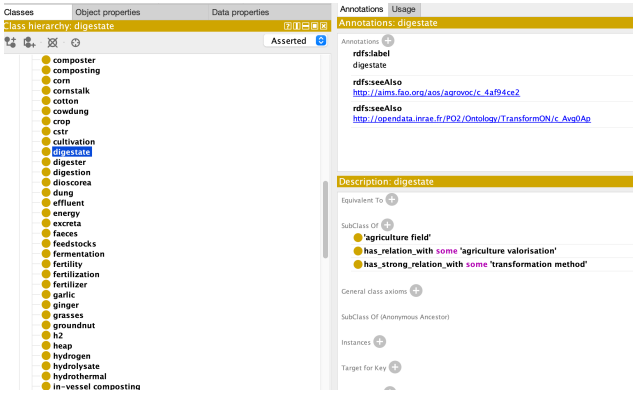


FIGURE 4 – Concept générique décrit dans les ontologies d'AGROVOC et PO2, et ayant des liens avec TM et AV

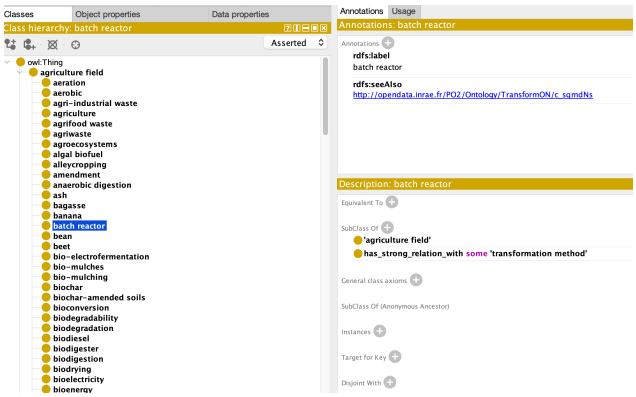


FIGURE 5 – Concept spécifique décrit uniquement dans l'ontologie PO2 et ayant un lien fort avec TM

4.3 Ontologie générée

L'ontologie générée est le résultat d'une génération automatique de classes et de relations, ainsi qu'un alignement ontologique. Elle comporte quatre classes parentes (voir Section 3.5.1) dont l'une contient 204 sous-classes qui sont les lexiques thématiques. Elle comporte également deux types de relations entre les classes : la relation hiérarchique et les propriétés d'objets qui sont au nombre de deux. Les figures 4, 5 et 6 illustrent ces différents types de classes et de relations.

TABLE 4 – Couverture d'alignement ontologique

Ressources	AGROVOC	PO2	AGOVOC & PO2	NA
Taux (%)	25	5	23	48

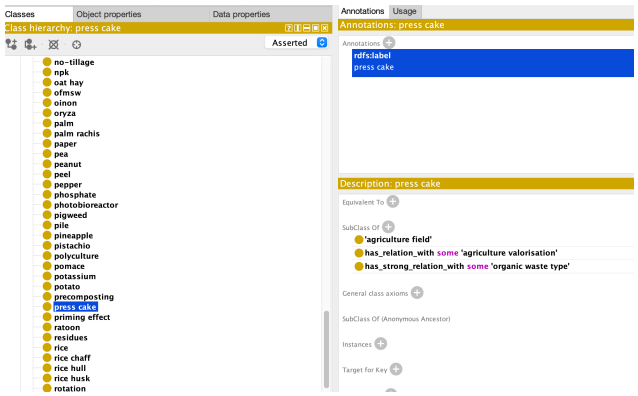


FIGURE 6 – Concept spécifique non décrit ni dans l'ontologie AGROVOC ni dans PO2, ayant des liens avec OWT et AV

Le Tableau 4 présente les taux de couverture d'alignement avec l'ontologie de AGROVOC et celle de PO2. Nous pouvons voir que 25% des lexiques sont décrits uniquement dans AGOVOC, et 23% sont décrits à la fois dans AGOVOC et dans PO2. Les lexiques que l'on ne trouve que dans PO2 représentent 5%. Même si PO2 est spécifique à la transformation des résidus, le taux de couverture de l'alignement avec cette ontologie n'est que d'environ 28% (23+5).

Nous avons ensuite vérifié manuellement les 48% restants afin de déterminer s'il existait des termes très spécifiques, c'est-à-dire des termes sans aucune correspondance, et des termes avec des correspondances, mais nécessitant une étude beaucoup plus approfondie pour choisir leurs correspondances parmi une liste de candidats. Nous avons alors constaté que 29,4% (60/204) de l'ensemble de nos termes étaient très spécifiques (ex. : *biodrying*, *in-vessel composting*, *vermifiltration*), ce qui n'est pas négligeable. La découverte de cette spécificité de certains lexiques est intéressante car elle permet d'une part de découvrir des lexiques spécifiques au domaine qui sont extraits d'articles scientifiques mais qui ne sont pas présents dans les ontologies connues, et d'autre part de décrire ces lexiques via notre ontologie et de rendre cette ontologie ouverte. Outre la découverte de lexiques spécifiques, la spécificité de notre ontologie réside dans la détection des liens sémantiques qu'un lexique peut avoir avec OTW, TM et AV, qui est importante dans le domaine. Sur ces 48%, 29,4% sont très spécifiques et les 18,6% restants présentent des correspondances lors de l'alignement, mais nécessitent une étude plus approfondie pour déterminer la correspondance. Cela peut concerner la diversité des noms (qui peuvent avoir la même racine ou être complètement différents), des groupes de mots plus spécifiques contenant le lexique en question (ex. : *nitrogen* correspond au concept *nitrogen compounds* d'AGROVOC), des abréviations (ex. : *h2* correspond au concept *hydrogen* d'AGROVOC), des nominalisations (ex. : *no-till* correspond au concept *zero-tillage* d'AGROVOC), etc. Même si nous avons utilisé des variantes pour la correspondance, nous avons tout de même manqué 18% des correspondances qui ne sont pas nécessairement liées à la syntaxe, mais au contexte. Notons que l'approche d'alignement ontologique adoptée dans ce travail pourra être étendue pour détecter les correspondances fondées sur la similarité contextuelle, la synonymie, etc. (ex. : la correspondance entre *pigweed* et *Portulaca oleracea*). Afin de rendre l'ontologie accessible au public et à la communauté, de manière à ce qu'elle puisse être réutilisée ou évaluée, nous prévoyons de la déposer dans *AgroPortal*[9], qui est le foyer des ontologies et des artefacts sémantiques dans l'agroalimentaire et les domaines connexes [10]. Afin d'enrichir les données et d'élargir l'accès aux techniques et aux concepts que nous découvrons grâce à notre approche, nous prévoyons également de procéder à une approche de matching des données avec les données ouvertes interconnectées (LOD : Linked Open Data), telle que Wikidata.

Comme indiqué dans l'introduction, l'un des objectifs de notre projet est de mettre en relation les différentes parties prenantes, et de leur fournir des informations sur les approches et techniques existantes en matière de transformations des déchets organiques ; ce cadre de normalisation de concepts ne constitue donc qu'une seule étape certe essentielle vers la future de représentation et de centralisation des données. Il convient de noter que dans ce travail la représentation des concepts thématiques et celle des concepts spatiaux sont traités séparément, mais qu'elles seront par la suite utilisées conjointement pour la contextualisation et la représentation du corpus d'articles scientifiques, à l'aide, par exemple, des graphes de connaissances. Toutes les étapes de notre approche (voir Figure 1) sont génériques et peuvent s'appliquer à tout type d'informations textuelles sur le traitement des déchets organiques, partout dans le monde. La seule condition est que les données soient en français.

5 Conclusion

Dans le but de promouvoir des approches de transformation des déchets organiques dans les pays du Sud, nous avons proposé une approche permettant d'extraire des lexiques spécifiques à ce domaine et des entités spatiales à partir d'articles scientifiques, une approche permettant de classer les lexiques par thème, et une approche permettant de générer et d'aligner automatiquement des ontologies à des fins de normalisation et de réutilisation. Des techniques de traitement du langage naturel et des approches d'ingénierie ontologique ont été utilisées. Un des avantages de l'approche proposée est qu'elle permet d'améliorer l'interopérabilité thématique d'un nouveau lexique ou d'un lexique peu connu avec des ressources du domaine. Elle permet également de contextualiser spatialement les lexiques par la prise en compte de l'information spatiale à travers des gazettiers. Ce travail présente des limites en termes d'identification des termes spécifiques dans ce domaine, ce qui nécessite des travaux supplémentaires. Nous prévoyons également de déposer l'ontologie dans *AgroPortal*. Nous prévoyons de mettre en œuvre une approche qui tient compte

9. https ://agroportal.eu

de la sémantique pour l'alignement ontologique, ainsi que d'effectuer des analyses sémantiques susceptibles d'intéresser les parties prenantes.

Références

[1] Hussein I Abdel-Shafy and Mona SM Mansour. Solid waste issue : Sources, composition, disposal, recycling, and valorization. *Egyptian journal of petroleum*, 27(4) :1275–1290, 2018.

[2] Explosion AI. spacy : Industrial-strength natural language processing in python, 2023.

[3] Foteini Barla, Filopoimin Lykokanellos, and Antonis C Kokossis. Discovering valorisation paths in waste biorefineries using an ontology engineering approach. In *Computer Aided Chemical Engineering*, volume 38, pages 2079–2084. Elsevier, 2016.

[4] Simona Bisiani, Agnes Gulyas, and Bahareh Heravi. Towards efficient and accessible geoparsing of u.k. local media : A benchmark dataset and llm-based approach. *Computational Humanities Research*, 1 :e10, 2025.

[5] Caterina Caracciolo, Armando Stellato, Ahsan Morshed, Gudrun Johannsen, Sachit Rajbhandari, Yves Jaques, and Johannes Keizer. The agrovoc linked dataset. *Semantic Web*, 4(3) :341–348, 2013.

[6] Pablo Castillo García, María José Fernández-Rodríguez, Rafael Borja, Juan Manuel Mancilla-Leytón, and David De la Lama-Calvente. Research trends in the recovery of by-products from organic waste treated by anaerobic digestion : a 30-year bibliometric analysis. *Fermentation*, 10(9) :446, 2024.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*, 2019.

[8] GeoNames. Geonames geographical database. `https://www.geonames.org/`, n.d. Accessed : 2026-02-16.

[9] Christian Jacquemin. FastR : A unification-based front-end to automatic indexing. In *Intelligent Multimedia Information Retrieval Systems and Management*, volume 1 of *RIAO '94*, pages 34–47, Paris, France, 1994. Centre de Hautes Études Internationales d'Informatique Documentaire.

[10] Clement Jonquet, Anne Toulet, Biswanath Dutta, and Vincent Emonet. Harnessing the power of unified metadata in an ontology repository : the case of agroportal. *Journal on Data Semantics*, 7(4) :191–221, 2018.

[11] Bach Thanh Le, Rose Dieng-Kuntz, and Fabien Gandon. On ontology matching problems. *ICEIS (4)*, pages 236–243, 2004.

[12] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. Biomedical term extraction : overview and a new methodology. *Inf. Retr.*, 19(1–2) :59–99, April 2016.

[13] Leticia Sarmento dos Muchangos. Mapping the circular economy concept and the global south. *Circular Economy and Sustainability*, 2(1) :71–90, 2022.

[14] Maureen Chiebonam Okibe, Michael Short, Franjo Cecelja, and Madeleine Bussemaker. Ontology modelling for valorisation of sugarcane bagasse. In *Computer Aided Chemical Engineering*, volume 52, pages 3363–3368. Elsevier, 2023.

[15] Christiane Rakotomalala, Jean-Marie Paillat, Frédéric Feder, Angel Avadi, Laurent Thuriès, Marie-Liesse Vermeire, Jean-Michel Médoc, Tom Wassenaar, Caroline Hottelart, Lilou Kieffer, et al. A lexicon obtained and validated by a data-driven approach for organic residues valorization in emerging and developing countries. *Frontiers in Artificial Intelligence*, 8 :1557137, 2025.

[16] Christiane Rakotomalala, Jean-Marie Paillat, Frédéric Feder, Angel Avadi, Laurent Thuriès, Marie-Liesse Vermeire, Jean-Michel Médoc, Tom Wassenaar, Caroline Hottelart, Lilou Kieffer, Elisa Ndjie, Mathieu Picart, Jorel Tchamgoue, Alvin Tulle, Laurine Valade, Annie Boyer, Marie-Christine Duchamp, and Mathieu Roche. A lexicon for organic residues valorization in emerging and developing countries, CIRAD Dataverse, 10.18167/DVN1/HNZZSI, 2023.

[17] Mathieu Roche, Agneta Lindsten, Tomas Lundén, and Thierry Helmer. Leap4fnssa lexicon : Towards a new dataset of keywords dealing with food security. *Data in Brief*, 45 :108680, 2022.

[18] Nikolaos Trokanas, Franjo Cecelja, and Tara Raafat. Semantic input/output matching for waste processing in industrial symbiosis. *Computers & Chemical Engineering*, 66 :259–268, 2014.

[19] Magalie Weber, Patrice Buche, Liliana Ibanescu, Stéphane Dervaux, Hervé Guillemin, Julien Cufi, Michel Visalli, Elisabeth Guichard, and Caroline Pénicaud. Po2/transformon, an ontology for data integration on food, feed, bioproducts and biowaste engineering. *npj Science of Food*, 7(1) :47, 2023.

[20] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. GLiNER : Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, pages 5364–5376, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[21] S. Zheng. Spatialwebagent : Leveraging large language models for spatial entity extraction. In *Proceedings of the ACL 2025 Demo Track*, pages 1–15, 2025.