KEOPS-CTS : Knowledge ExtractOr Pipeline System pour l'analyse de Champs Thématiques Stratégiques

S. Valentin[1,2], T. Helmer[3], X. Rouvière[3], M. Roche[1,2]

[1] CIRAD, UMR TETIS, 34398 Montpellier, France
[2] TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France
[3] CIRAD, DSI, 34398 Montpellier, France

sarah.valentin@cirad.fr

Résumé

Les outils de gestion des connaissances visent à faciliter les processus de collecte et d'organisation des connaissances afin de rendre ces connaissances disponibles dans une base partagée. Nous présentons la plateforme KEOPS-CTS dédiée à l'extraction et la gestion de connaissances à partir de données textuelles produites par un organisme de recherche qui traite les problématiques en agriculture appliquées aux pays du Sud. La démarche de KEOPS est guidée par les informations sémantiques (CTS - Champ Thématique Stratégique) contenues dans les textes hétérogènes intégrés à la plateforme. Nous proposons une évaluation de l'expansion lexicale afin d'améliorer l'analyse des documents relatifs à l'agroécologie.

Mots-clés

Système de gestion des connaissances, Fouille de texte, Expansion du vocabulaire, Plongements lexicaux

Abstract

Knowledge management tools aim to facilitate the process of collecting and organising knowledge to make it available in a shared database. We present the KEOPS-CTS platform, dedicated to the extraction and management of knowledge from textual data produced by a research organization addressing agricultural issues in the Global South. The KEOPS approach is guided by the semantic information (CTS - Strategic Thematic Field) contained in the heterogeneous texts integrated into the platform. We propose an evaluation of lexical expansion to improve the analysis of documents related to agroecology.

Keywords

Knowledge management system, Text mining, Vocabulary expansion, Word embeddings

1 Introduction

Le processus de gestion de connaissances à partir de données (*Knowledge Management*) implique généralement la succession de plusieurs étapes, parmi lesquelles le nettoyage, l'indexation, la sélection et la transformation des données avec l'utilisation éventuelle de modèles. Á ces étapes peut s'ajouter la mise en forme des résultats via des choix de visualisation adaptés.

Quel que soit le domaine d'application, une part importante de l'information disponible est stockée sous forme de texte, dans des documents provenant de sources hétérogènes telles que des documents officiels, institutionnels, des contenus de site web, des communications, etc. Les données textuelles sont dites "non-structurées" et nécessitent des techniques de recherche d'information et d'analyse adaptées fondées sur la fouille de texte et le traitement automatique du langage naturel (TALN). Par exemple, [1] utilise une approche de classification non supervisées afin de regrouper des documents d'ingénierie sur la base de leur proximité sémantique. [14] compare des termes extraits par une mesure de pondération classique avec des termes identifiés suite à un clustering automatique dans un processus de classification automatique de la polarité de documents d'évaluation de projet. D'autres travaux se sont intéressés à la classification de connaissances textuelles à partir documents normatifs [6], dans une approche supervisée reposant sur des modèles de langue pré-entrainés de type *Transformers*.

Parallèlement aux travaux de recherche, les approches d'analyse de données textuelles sont de plus en plus intégrées aux outils de gestion des connaissances (Knowledge Management Systems, ou KMS). Ces outils ont pour but de soutenir "l'un des trois processus fondamentaux de gestion des connaissances : la génération, la codification et le transfert de connaissances." [3]. Par exemple, TyDI (Terminology Design Interface) est une plateforme collaborative pour la validation manuelle et la structuration de termes à partir de terminologies existantes ou de termes extraits automatiquement à l'aide d'outils dédiés [11]. D'autres outils comme NooJ [13] utilisent des approches linguistiques pour construire et gérer des dictionnaires et des grammaires. NooJ intègre plusieurs méthodes de traitement du langage naturel, comme les approches de reconnaissance des entités nommées. D'autres plateformes intègrent des composants d'exploration de texte comme CorTexT [2]. CorTexT permet l'extraction d'entités nommées et des approches avancées d'exploration de texte (par exemple, la modélisation de sujets, l'intégration de termes, etc.) sont intégrées dans

cette plateforme.

La création d'outils de gestion des connaissances a un tropisme historique dans le domaine des pratiques organisationnelles en entreprises [3, 4]. Cependant, la création et l'utilisation efficace de l'information et du savoir sont aussi des besoins clés dans le domaine de la recherche et de la gestion de projets scientifiques. La plateforme AReS (*Agricultural Research e-Seeker*) est une plateforme qui permet d'explorer et d'extraire du contenu à partir de dépôts d'informations et de données textuelles liés au Groupe consultatif pour la recherche agricole internationale (CGIAR) et à ses partenaires [1]. Cet outil est conçu pour aider à rendre les connaissances du CGIAR trouvables, accessibles, interopérables et réutilisables et propose une indexation avec Agrovoc, un thésaurus dédié au domaine agricole [2].

KEOPS (Knowledge ExtractOr Pipeline System) est une plateforme qui applique diverses méthodes d'indexation et de classification à des données textuelles provenant de bases de données, de documents ou de pages web [9]. Une caractéristique de KEOPS est de guider l'indexation, l'analyse et la visualisation des informations et connaissances produites selon un angle sémantique. Ce dernier s'appuie sur un vocabulaire contrôlé. Par exemple, dans le cadre du projet LEAP4FNSSA [3], les documents ont été analysés selon un lexique lié à la sécurité alimentaire [12].

En sortie, KEOPS combine les résultats de classification et d'indexation pour générer des connaissances sur chaque texte et groupe de textes.

Dans cet article, nous décrivons l'adaptation de l'outil KEOPS à la gestion de documents sous le prisme des Champs Thématiques Stratégique (CTS) d'un organisme de recherche, le Cirad (section 2). Nous présentons ensuite notre méthodologie sur l'expansion d'un vocabulaire relatif à l'agroécologie (section 3), les résultats du cas d'étude proposé (section 4) et une discussion.

2 KEOPS-CTS

2.1 Données

L'objectif de l'outil KEOPS-CTS est de permettre la collecte, l'indexation et l'analyse de données textuelles issues de différentes sources et caractérisant différentes étapes du cycle de vie de l'activité de recherche du Centre de coopération internationale en recherche agronomique pour le développement (Cirad) : l'orientation scientifique, les activités en cours, et les productions scientifiques associées (Figure 1). Les données et leurs sources sont détaillées ci-après :

2.1.1 Orientation scientifique

Cet axe est représenté par (1) les profils de poste, qui renseignent sur la manière dont les compétences techniques et scientifiques sont utilisées dans les domaines couverts par les Champs Thématiques Stratégique (CTS), (2) les lettres pluriannuelles d'objectif (LPO) qui explicitent la politique

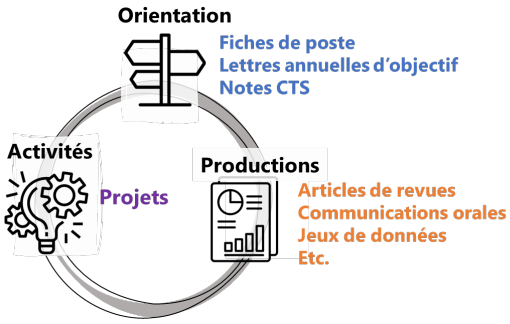

FIGURE 1 – Données textuelles intégrées dans l'outil KEOPS-CTS.

scientifique et partenariale des unités de recherche, en cohérence avec la stratégie du Cirad et (3) les notes CTS, qui font un état des lieux des recherches conduites au Cirad dans chacun des CTS, les enjeux spécifiques et les fronts de science sur lesquels le Cirad souhaite être visible avec ses partenaires.

2.1.2 Activités scientifiques

Le contenu des activités est analysé sous le prisme des résumés des projets dans lesquels le Cirad est impliqué et recensés sur la plateforme CORDIS [4]), principale source des projets financés par les programmes cadres de l'Union Européenne.

2.1.3 Productions scientifiques

Enfin, nous avons extrait de l'archive ouverte du Cirad, (Agritrop [5]), l'ensemble des résumés correspondants aux productions suivantes : articles de revues scientifiques avec ou sans comité de lecture, communications avec actes, ouvrages, thèses et jeux de données.

Au moment de l'extraction, la base de données textuelles finale contient 18721 documents (Tableau 1).

Type de source	Nombre de documents	Nombre moyen de termes
Orientation scientifique		
Fiches de postes	1776	290
Lettres d'objectifs	1047	350
Notes CTS	14	6096
Activités		
Projets	121	439
Productions		
Articles de revue	9386	246
Ouvrages	2932	233
Communications	2383	244
Thèses	543	247
Jeux de données	519	177

TABLE 1 – Nombre de documents par type de source au 1/03/2024

1. https://cgspace.cgiar.org/explorer/
2. https://agrovoc.fao.org/browse/agrovoc/en/
3. Long-term Europe-Africa Research and Innovation Partnership for Food and Nutrition Security and Sustainable Agriculture
4. https ://cordis.europa.eu/
5. https ://agritrop.cirad.fr/

2.2 Classification

Un module de classification est intégré dans l'outil KEOPS-CTS afin de déterminer automatiquement le type de source de chaque document. Cette classification repose sur une approche supervisée : plusieurs familles de classifieurs (e.g. Random Forest, Multilayer Perceptron) sont entraînées sur un jeu de données annotées afin de prédire le type de source de chaque nouveau document [9].

2.3 Indexation

L'indexation dans KEOPS CTS repose sur trois types d'approches :

- L'indexation thématique, réalisée à partir de vocabulaires thématiques construits par des experts (termes sources et leurs synonymes)
- L'indexation thématique réalisée à partir de thésaurus spécialisés tel que Agrovoc [6].
- L'indexation par une terminologie acquise automatiquement à l'aide de BioTex [8] ;
- L'indexation automatique par un ensemble d'entités nommées (e.g. lieux, organisations, etc.) extraites par un modèle pré-entrainé issu de la librairie SpaCy [7].

Chaque document est indexé avec l'ensemble de ces approches (Figure 2), ce qui permet de réaliser des requêtes combinant informations thématiques et informations transversales (e.g. localisations).

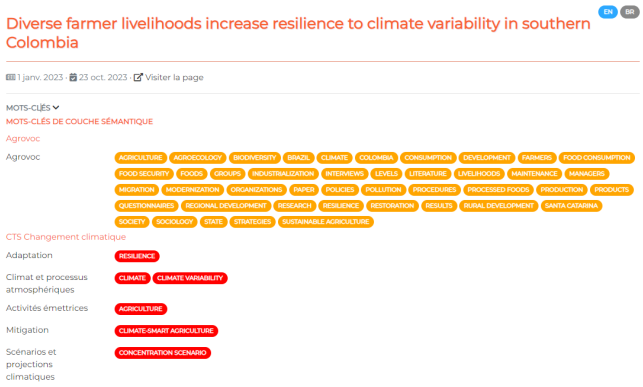

FIGURE 2 – Interface de KEOPS-CTS montrant la liste des vocabulaires d'indexation et termes associés.

3 Constitution de vocabulaires thématiques

Les lexiques (ou vocabulaires) thématiques sont des entrées indispensables au processus d'extraction de connaissances. Ils constituent l'angle de vue expert à partir duquel les données textuelles vont être indexées et analysées. Les termes, qui constituent un vocabulaire donné, visent à refléter de façon la plus exhaustive possible les concepts associés à une thématique. Les lexiques correspondant aux différents CTS sont initialement construits grâce à une méthode itérative combinant avis d'expert et méthodes d'extraction automatique [12, 5].

Dans les sections suivantes, nous décrivons les approches proposées pour étendre automatiquement les termes source d'un vocabulaire donné, en les évaluant à travers un lexique dédié à l'agroécologie.

3.1 Expansion du lexique

La découverte de synonymes à partir d'un corpus massif est une tâche indispensable pour la découverte automatique de connaissances : elle permet d'améliorer les tâches d'indexation et de recherche d'information. Pour un terme donné, ses synonymes font référence aux termes qui peuvent être utilisés de manière interchangeable dans certains contextes. Nous avons comparé deux méthodes pour l'expansion de vocabulaires source, i.e. le plongement lexical (*word embedding*) et une approche fondée sur un modèle de langue génératif (Figure 3).

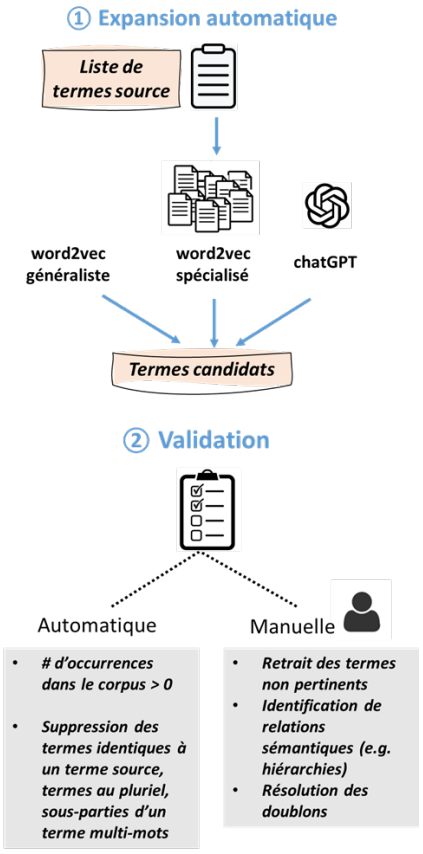

FIGURE 3 – Approche globale d'expansion du vocabulaire source combinant approches automatiques et validation d'expert.

Ci-après, l'expression "terme source" désigne un terme issu du vocabulaire constitué par les experts. Il s'agit des termes pour lesquels nous souhaitons obtenir des synonymes. Ils incluent des termes formés d'un seul mot (e.g. *agroforestry*) et des termes formés de plusieurs mots, ou "termes multi-mots" (e.g. *family farming*). Les approches d'expansion ont

6. https://agrovoc.fao.org/browse/agrovoc/en/
7. https://spacy.io/

été évaluées manuellement sur la base de trois termes par terme source.

3.1.1 Modèles de plongement lexical

Les modèles de word embedding ont été introduits par Mikolov et al. en 2013 [10]. Cette approche repose sur l'entraînement de modèles à partir de large corpus afin de traduire l'ensemble des termes (vocabulaire) en une représentation numérique de haute dimension. Cette représentation numérique des termes peut être utilisée pour comparer différents termes et trouver des termes similaires sur le plan sémantique. Nous avons comparé deux stratégies d'obtention d'*embeddings* :

- l'utilisation d'un modèle de word embedding pré-entrainé à partir d'un large corpus (modèle w2vG) et disponible sur HuggingFace [8]
- l'entraînement de modèles spécialisés à partir de notre corpus d'application (modèles w2vS). Les paramètres d'entrainement correspondent aux paramètres par défaut du modèle Word2vec implantés dans la librairie python Gensim, en faisant varier le type d'architecture (skip-gram et CBOW).

Les modèles w2vS skip-gram et CBOW ont été entrainés à partir de l'ensemble des documents de la base de KEOPS-CTS. Le modèle ne produit de représentation vectorielle que pour les termes initialement présents dans le corpus d'apprentissage. Afin de générer des représentations des termes multi-mots du vocabulaire source lors de l'entrainement des modèles w2vS, ces termes ont donc préalablement été détectées et concaténées avec le séparateur '_' dans le corpus d'entrainement. Le corpus a été converti en minuscule et les caractères non alphanumériques ont été retirés.

3.1.2 Modèle de langue génératif

Notre seconde approche a consisté à utiliser la capacité générative du modèle ChatGPT-3 afin de générer des synonymes. Contrairement à l'entrainement d'un modèle de plongement de termes, l'utilisation de ChatGPT ne nécessite pas de corpus d'apprentissage et le vocabulaire n'est pas restreint. Le prompt utilisé a été le suivant : "*Pour chacun de ces termes ou termes constitués de plusieurs mots, proposez au maximum trois synonymes. Les synonymes doivent être sémantiquement et grammaticalement corrects.*"

3.2 Validation

L'étape de validation manuelle consistait initialement à valider les synonymes pertinents pour chaque terme source. Or, lors de l'évaluation préliminaire, nous avons identifié un second niveau de pertinence, à savoir des termes ne correspondant pas à des synonymes, mais pouvant être d'intérêt dans le cadre de l'extension d'un vocabulaire. Trois cas ont été distingués : les termes correspondant à un concept hiérarchique supérieur, ou hyperonymes (e.g. *fertilizer* pour *green manure*), à un concept hiérarchique inférieur, ou hyponymes (e.g. *bean* pour *legume*) ou à un autre concept pertinent sans relation hiérarchique clairement identifiable (*pastoralism* pour *family farming*).

8. https ://huggingface.co/fse/word2vec-google-news-300

Pour les termes non pertinents, nous avons distingué les termes non pertinents pour le domaine d'application ou (e.g. *darling* proposé comme synonyme de *honey*) et termes ou expressions incorrects (orthographe incorrecte, sous-partie d'un terme multi-mot).

3.3 Evaluation

Au-delà de la pertinences des synonymes, nous souhaitons en évaluer la capacité d'expansion. Pour cela, nous avons défini le coefficient d'expansion (CE), à l'échelle de l'occurrence et à l'échelle du document :

- Le coefficient d'expansion à l'échelle de l'occurrence (*CEocc*) correspond au coefficient multiplicateur entre le nombre d'occurrences d'un terme source dans le corpus, et la somme du nombre d'occurrences du même terme et de son synonyme.
- Le coefficient d'expansion à l'échelle du document (*CEdoc*) correspond au coefficient multiplicateur entre (i) le nombre de documents dans lesquels apparait un terme, (ii) le nombre de documents dans lesquels apparaissent un terme ou son synonyme.

Pour une méthode d'expansion donnée, son index d'expansion est calculé en faisant la moyenne de tous les synonymes générés par cette méthode. Les termes au singulier et leur version au pluriel sont considérés. Pour pouvoir calculer ce coefficient dans le cas où le terme source n'apparait pas dans le corpus mais où le synonyme proposé est détecté, l'occurrence du terme a été arbitrairement définie à 1.

4 Résultats

Dans cette section, nous présentons l'évaluation des différentes approches d'expansion à partir d'un vocabulaire dédié à l'agroécologie. Ce vocabulaire source est dérivé d'un lexique construit par avis d'experts [5]. Il contient 213 termes source, parmi lesquels 26 termes simples et 187 termes multi-mots.

4.1 Vocabulaires des modèles

Les modèles de plongements lexicaux se basent sur des vocabulaires fixes définis par leur corpus d'entraînement et les étapes de prétraitement appliquées, telles que la suppression des nombres et des caractères spéciaux, ou la concaténation des termes multi-mots. Par conséquent, leur taux de couverture d'un vocabulaire spécifique (ensemble des termes d'un corpus) varie. Concernant le vocabulaire lié à l'agroécologie, le modèle w2vG obtient les résultats les moins satisfaisants, particulièrement pour les termes multi-mots. Ayant été entraîné sur un corpus spécialisé, le modèle w2vS couvre la quasi-totalité des termes simples, mais seulement 45% des termes multi-mots (Tableau 2).

4.2 Validation des termes issus des méthodes d'expansion

Le nombre total de synonymes pertinents après validation manuelle était de 25 pour w2vG, 12 pour w2vS (cbow), 13 pour w2vS (skip) et 433 pour ChatGPT. ChatGPT et w2vG obtiennent les meilleures proportions de synonymes pertinents (71% et 39.1%, respectivement) (Tableau 3).

Méthode d'expansion	Termes simples	Termes multi-mots
w2vG	69.3%	3.7%
w2vS	92.3%	45%
ChatGPT	100%	100%

TABLE 2 – Taux de couverture des différentes méthodes d'expansion

La proportion de synonymes parmi les termes proposés par les modèles d'embedding spécialisés (w2vS) sont très faibles (autour de 4%). Cependant, ces modèles proposent des concepts pertinents de type hyperonyme, hyponyme et autres concepts associés que ne génère pas ChatGPT.

	w2vG	w2vS (cbow)	w2vS (skip)	ChatGPT
Pertinent - synonyme	39.1%	3.8%	4.1%	71%
Pertinent - autre				
Hyperonyme	0%	2.6%	1.9%	1.1%
Hyponyme	7.8%	3.2%	1.6%	1.3%
Autre concept	17.2%	12.8%	15.9%	0.7%
Non pertinent :				
Terme non pertinent	14.1%	75.0%	73.2%	25.9%
Pluriel	12.5%	1.9%	2.2%	0%
Mauvaise orthographe	6.2%	0%	0%	0%
Sous-partie	3.1%	0.6%	1.0%	0%

TABLE 3 – Évaluation de la pertinence des termes issus des différentes méthodes d'expansion. Pour chaque méthode, les proportions représentent le nombre de termes de chaque catégorie, par rapport au nombre total de termes obtenus par cette méthode.

La répartition des termes source en fonction du nombre de synonymes pertinents générés par les différentes approches est résumée dans le Tableau 4. Les modèles de plongement de mots étudiés sont très peu performants du point de vue de la synonymie. La prise en compte de l'occurrence des synonymes dans le corpus impacte très négativement la performance de ChatGPT : 52% des termes source, les synonymes proposés sont non pertinents ou pertinents mais absents du corpus. En effet, bien que généralement correctes d'un point de vue grammatical, les propositions de ChatGPT sont parfois des constructions terminologiques peu susceptibles d'être utilisées dans un corpus spécialisé (par exemple, *eco-friendly fertilizer* pour *biofertilizer*). Les modèles w2vS ayant été entrainés sur le corpus, les termes proposés y apparaissent nécessairement.

4.3 Evaluation de l'expansion

Les coefficients d'expansion de ChatGPT sont significativement supérieurs à ceux des modèles d'embeddings : le nombre de détections d'occurrences est multiplié en moyenne par 12.6. Notamment, pour 24 termes source n'apparaissant pas dans le corpus, ChatGPT a généré des synonymes permettant de détecter le terme initial (e.g. *indigenous species* permettant de détecter le terme *native breed*). Ce comportement offre un gain conséquent en termes d'indexation.

	w2vG	w2vS (cbow)	w2vS (skip)	ChatGPT
Synonymes pertinents				
0	93.3%	93.9%	92.8%	1.6%
1	1.1%	5.5%	7.2%	16.6%
2	3.9%	0.6%	0%	22.7%
3	1.7%	0%	0%	59.1%
Synonymes pertinents et présents				
0	96.0%	93.9%	92.8%	52.0%
1	1.7%	5.5%	7.2%	26.5%
2	1.7%	0.6%	0%	16.0%
3	0.6%	0%	0%	5.5%

TABLE 4 – Proportion de termes source en fonction du nombre de synonymes pertinents obtenus par les différents modèles et du nombre de synonymes pertinents présents dans le corpus.

	w2vG	w2vS0	w2vS1	ChatGPT
CEocc	1.13	2.30	1.80	12.6
CEdoc	1.15	2.44	1.86	7.5
Nb termes source	12	11	13	178

TABLE 5 – Comparaison du coefficient d'expansion entre les différents modèles, à l'échelle du nombre d'occurrences (*CEocc*) et du nombre de documents (*CEdoc*).

5 Discussion

Dans ces travaux préliminaires, nous avons comparé deux familles d'approches pour l'expansion de vocabulaire sur la thématique de l'agroécologie, i.e. le plongement lexical et un modèle de langue génératif. Une différence fondamentale entre ces deux approches repose sur la définition de la tâche : la proximité vectorielle dans l'espace d'un modèle de word embedding ne correspond pas nécessaire à une relation de synonymie. Au contraire, la nature d'une tâche peut être explicitement définie lors du prompt associé à modèle génératif, ce qui assure une homogénéité des résultats produits. ChatGPT s'est montré particulièrement performant pour l'expansion de termes multi-mots, tâche pour laquelle les modèles de plongements lexicaux nécessitent des étapes de pre-processing adaptées et offrent des performances moindres. Les modèles d'embeddings permettent d'identifier des termes issus de relations hiérar-

chiques diverses et susceptibles d'enrichir le vocabulaire source. Lorsque les modèles sont appris à partir d'un corpus spécialisé, ils peuvent participer à l'identification de concepts pertinents non identifiés par les experts et participer à l'enrichissement d'une hiérarchie. L'évaluation de prompts dédiés à la recherche de termes issus de différents types de hiérarchies est cependant nécessaire afin de comparer les deux approches vis-à-vis de cette tâche. De plus, l'adaptation de modèles de langues sur notre corpus thématique (*fine-tuning*) pourrait significativement améliorer les performances des tâches d'extraction de synonymes et autres types de liens sémantiques [7].

6 Conclusion

Nous présentons KEOPS-CTS un système de gestion des connaissances dédié aux données textuelles produites par des activités de recherche. L'intégration des récentes avancées en traitement automatique de la langue et Intelligence Artificielle ne sont encore que peu incorporées dans les outils open-source. Nous proposons une première contribution sur l'expansion lexicale en comparant des modèles de plongements lexicaux et un modèle de langue génératif, ChatGPT, et en montrons leur complémentarité.

Les futurs travaux consisteront à intégrer ces approches et à évaluer le résultat de ces expansions à travers les différents axes d'analyse, en particulier selon les types de documents (orientation, activités et productions scientifiques).

Remerciements

Nous remercions Julien Rabatel pour le développement de KEOPS, les personnes de la Délégation à l'information scientifique et à la science ouverte et le service des ressources humaines du Cirad pour l'extraction des données. Ces travaux menés dans le cadre du projet CEA-First ont reçu le financement de l'Union Européenne - Programme HORIZON - Grant Agreement No. 101136771. Les données sources en agroécologie ont été acquises dans le cadre du projet ASSET (AFD, Union Européenne, FFEM).

Références

[1] Ivar Örn Arnarsson, Otto Frost, Emil Gustavsson, Mats Jirstrand, and Johan Malmqvist. Natural language processing methods for knowledge management—Applying document clustering for fast search and grouping of engineering documents. *Concurrent Engineering*, 29(2) :142–152, June 2021.

[2] Philippe Breucker, Jean-Philippe Cointet, Alexandre Hannud Abdo, Guillaume Orsal, Constance de Quatrebarbes, Tam-Kien Duong, Cristian Martinez, Juan Pablo Ospina Delgado, Luis Daniel Medina Zuluaga, Diego Fernando Gómez Peña, Tatiana Andrea Sánchez Castaño, Joenio Marques da Costa, Hajar Laglil, Lionel Villard, and Marc Barbier. Cortext manager. `https://docs.cortext.net`, October 2016.

[3] Thomas Davenport and Laurence Prusak. Working knowledge : How organizations manage what they know. *Ubiquity*, 1, January 1998.

[4] Rodrigo Valio Domínguez Gonzalez and Manoel Fernando Martins. Knowledge Management : an Analysis From the Organizational Development. *Journal of technology management & innovation*, 9(1) :131–147, April 2014.

[5] Thierry Helmer, Mathieu Roche, Pierre Martin, François Enten, Lucie Reynaud, Marie-Christine Lebret, Estelle Bienabe, Melanie Blanchard, Albrecht Ehrensperger, Ricardo Hernandez, and Germain Priour. ASSET Theoretical Lexicon : An agroecology lexicon. `https://dataverse.cirad.fr/dataset.xhtml?persistentId=doi:10.18167/DVN1/TVN3AC`, January 2023.

[6] Gu Jianan, Ren Kehao, and Gao Binwei. Deep learning-based text knowledge classification for whole-process engineering consulting standards. *Journal of Engineering Research*, July 2023.

[7] Ehsan Latif and Xiaoming Zhai. Fine-tuning chatgpt for automatic scoring. *Computers and Education : Artificial Intelligence*, 6 :100210, 2024.

[8] Juan Antonio Lossio-Ventura, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. BIOTEX : A system for Biomedical Terminology Extraction, Ranking, and Validation. In *International Semantic Web Conference*, 2014.

[9] Pierre Martin, Thierry Helmer, Julien Rabatel, and Mathieu Roche. KEOPS : Knowledge ExtractOr Pipeline System. In *Research Challenges in Information Science*, volume 415, pages 561–567. 2021.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv :1301.3781 [cs]*, January 2013.

[11] Claire Nédellec, Wiktoria Golik, Sophie Aubin, and Robert Bossy. Building large lexicalized ontologies from text : A use case in automatic indexing of biotechnology patents. In *Knowledge Engineering and Management by the Masses*, pages 514–523, 2010.

[12] 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, December 2022.

[13] Max Silberztein and Agnès Tutin. NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE. *Apprentissage des langues et systèmes d'information et de communication (Alsic)*, (Vol. 8, n° 1) :123–134, December 2005.

[14] Nadeem Ur-Rahman and Jenny Harding. Textual data mining for industrial knowledge management and text classification : A business oriented approach. *Expert Systems with Applications*, 39(5) :4729–4739, April 2012.