Regards – Focus

Comment mobiliser des approches de fouille de textes et d'extraction de la terminologie dans un contexte pluridisciplinaire ?

Mathieu Roche ^{*}

Informatique, CIRAD, UMR TETIS (Université de Montpellier, AgroParisTech, CIRAD, CNRS, INRAE), Montpellier, France

Résumé – L'analyse des masses de données nécessite l'utilisation de méthodes mêlant harmonieusement différentes disciplines comme l'informatique, les mathématiques, les statistiques. L'ensemble de ces méthodes utiles pour traiter de telles données forme le socle de la « science des données ». Dans ce cadre, les approches de fouille de textes permettent de découvrir des connaissances utiles et nouvelles pour des experts issus généralement de différents domaines d'application (par exemple, veille épidémiologique, sécurité alimentaire, etc.). Cet article dresse un panorama de l'utilisation de méthodes de fouille de textes dans différents projets liés à l'agriculture et à la santé. Une démarche méthodologique générique est ensuite proposée et discutée.

Mots-clés : fouille de textes / terminologie / corpus / agriculture / santé

Abstract – How to integrate text-mining and terminology extraction approaches in a multidisciplinary context? The analysis of large amounts of data requires using methods that combine a range of fields such as computer science, mathematics, statistics, etc. All of these methods useful for data processing form the basis of “data science”. In this context, text mining approaches allow the discovery of new and useful knowledge for experts generally originating from different application areas (e.g., epidemiological surveillance, food security, etc.). This paper provides an overview of the use of text mining methods in various projects related to agriculture and health. A generic methodological approach is then proposed and discussed. This is based on three stages:

- (1) Data collection. Corpus acquisition from the Web can be done with queries on search engines or RSS feeds.
- (2) Extraction of terminology using text-mining approaches. Terminology is automatically extracted using different parameters of the BioTex tool (e.g., F-TFIDF-C and C-value measures) dealing with texts in English, French and Spanish.
- (3) Validation of terms with end-users and field experts based on different approaches (e.g., surveys, workshops, etc.).

Keywords: text-mining / terminology / corpus / agriculture / health

Introduction

Les masses de données textuelles aujourd'hui disponibles engendrent un problème difficile lié à leur traitement automatique. Des méthodes de fouille de textes (FT) peuvent en partie répondre à une telle problématique. Elles consistent à modéliser puis mettre en œuvre des méthodologies appliquées aux données textuelles afin d'en déterminer le sens et/ou de découvrir

des connaissances nouvelles. Les descripteurs linguistiques représentent le matériau de base afin d'associer une certaine sémantique aux documents. Ces descripteurs peuvent être des mots simples (par exemple, « irrigation ») mais aussi des termes composés (par exemple, « agriculture familiale »). Nous appellerons de tels descripteurs linguistiques des « termes ».

Outre une synthèse présentant de quelle manière les méthodes d'extraction de la terminologie ont été mises en place dans différents projets, cet article a trois autres objectifs. (i) Il propose un cadre générique et une

* Auteur correspondant : mathieu.roche@cirad.fr

démarche commune. En effet, les méthodes de fouille de textes mises en œuvre étaient jusqu'à présent déclinées selon les différents projets sans formalisation de la démarche de manière structurée, unifiée et générique. (ii) Il permet de partager des expériences spécifiques mises en place dans chacun des projets qui peuvent être appliquées de manière similaire ou proche à d'autres. (iii) Il a pour ambition de développer un regard croisé pluridisciplinaire à travers des projets en santé et agriculture (compétences thématiques) en y intégrant un volet informatique important (compétences méthodologiques en fouille de textes).

Les parties 2 et 3 décrivent respectivement l'état de l'art et des applications qui ont mobilisé des approches de fouille de textes. La démarche méthodologique générique associée à ces applications et quelques résultats et discussions sont présentés en parties 4 et 5.

État de l'art

Les approches de recherche d'information (RI) sont, en général, guidées par l'utilisateur *via* des recherches ciblées. Les processus de fouille de textes adoptent une philosophie différente car ils permettent de mettre en avant des connaissances nouvelles, pépites de connaissances dans le meilleur des cas, à partir de données textuelles (textes libres) en mobilisant et combinant des approches statistiques et linguistiques.

Dans ce contexte, les termes du domaine général (Kennedy, 2010; Vakkari, 2010) ou de spécialité (Turenne et Barbier, 2004; Bartol, 2009; Névélol *et al.*, 2014) constituent le matériau de base des connaissances du domaine. L'identification des termes dans les textes libres peut être guidée (a) par consensus avec les experts (Laporte *et al.*, 2012), (b) par les données qui nécessitent, par exemple, la mise en œuvre de méthodes de fouille de textes (Dobrov et Loukachevitch, 2011; Lossio-Ventura *et al.*, 2016). Comme dans les travaux de Batifol *et al.* (2020), la démarche présentée dans cet article est mixte en combinant une approche guidée par les données (fouille de textes) tout en s'appuyant sur les connaissances thématiques (connaissances expertes) tout ou long du processus mis en œuvre.

Les méthodes classiques d'extraction de la terminologie sont fondées sur des approches statistiques et/ou morphosyntaxiques. Le système TERMINO (David et Plante, 1990) est un outil précurseur qui s'appuie sur une analyse morphologique à base de règles pour extraire les termes nominaux (aussi appelés syntagmes nominaux). Les travaux de Smadja (1993) (approche XTRACT) s'appuient sur une approche statistique. XTRACT extrait, dans un premier temps, les syntagmes binaires situés dans une fenêtre de dix mots. Les syntagmes (groupes de mots) binaires sélectionnés sont ceux qui

dépassent d'une manière statistiquement significative la fréquence due au hasard. L'étape suivante consiste à extraire les groupes de mots contenant les syntagmes binaires trouvés à la précédente étape. ACABIT (Daille, 1994) effectue une analyse linguistique afin de transformer les syntagmes nominaux en termes binaires. Ces derniers sont ensuite triés selon des mesures d'association entre éléments composant les syntagmes. Les mesures d'association et les approches distributionnelles ont été étendues et adaptées pour extraire des termes spécialisés (Frantzi *et al.*, 2000) ou identifier des termes synonymes (Hazem et Daille, 2014).

De nombreux outils et plateformes¹ sont disponibles et peuvent être utilisés dans un cadre pluridisciplinaire. Certains ont des vocations d'annotation (par exemple, GLOZZ², BRAT³, etc.), d'autres intègrent des outils de fouille de textes et d'extraction de la terminologie. Par exemple, TyDI (Terminology Design Interface) (Nédellec *et al.*, 2010) est une plateforme collaborative pour la validation manuelle et la structuration de termes issus de terminologies existantes ou de termes extraits automatiquement à l'aide d'outils dédiés, à savoir BioYatea (Aubin et Hamon, 2006). D'autres outils ont des approches linguistiques fines comme l'environnement NooJ qui permet de construire et de gérer des dictionnaires et grammaires (Silberstein, 2015). NooJ a aussi été intégré à de nombreuses applications de traitement automatique du langage naturel (TALN), par exemple la reconnaissance d'entités nommées, la génération automatique de textes, l'informatique décisionnelle, etc. D'autres plateformes comme UNITEX⁴ s'appuient sur des dictionnaires et des grammaires pour effectuer notamment des concordances de termes à partir d'expressions régulières ou de graphes, appelés aussi grammaires locales (Paumier, 2020; Kyriacopoulou *et al.*, 2018). D'autres plateformes intègrent des composantes de fouille de textes à l'image de CorText⁵ dédiée aux sciences humaines et sociales (SHS) (Barbier et Cointet, 2012). Un des objectifs de cette plateforme est de produire des analyses issues des données textuelles. Par exemple, CorText permet d'extraire des entités nommées en y adossant une couche spatiale. Par ailleurs, des fonctions avancées de fouille de textes sont adossées à cette plateforme (*topic modeling*, *word embedding*, etc.). CorText permet également la visualisation des termes et des associations entre ces derniers en

¹ <http://explorationdecorpus.corpusecrits.huma-num.fr/outils-logiciels-corpus-ecrits/>.

² <http://www.glozz.org/>.

³ <https://brat.nlplab.org/>.

⁴ Développé par le Laboratoire d'automatique documentaire et linguistique (LADL), fondé en 1968 par Maurice Gross.

⁵ <https://www.cortext.net/>.

s'appuyant sur la notion de cooccurrences (Roche, 2020). Prospero⁶ est une autre plateforme largement utilisée pour l'analyse des données en SHS en proposant notamment des fonctionnalités de classification et de suivi de thématiques à partir de données textuelles (Chateauraynaud, 2003). Ces différentes plateformes permettent de visualiser les termes en contexte ce que n'autorisent pas nécessairement d'autres outils cependant très souples et circonscrits quant à leur utilisation (Lossio-Ventura *et al.*, 2016). Tous ces outils et plateformes peuvent réaliser des analyses à partir de textes de différents domaines, en particulier dans les domaines de l'agriculture (Malanski *et al.*, 2021) et de la santé (Gauld et Micoulaud-Franchi, 2020) qui sont étudiés dans cet article.

Applications de la fouille de textes dans des travaux pluridisciplinaires

Dans le cadre de ces travaux, nous résumons, dans cette partie, (i) les problématiques générales associées à des tâches dédiées et (ii) de quelle manière les approches de fouille de textes et d'extraction de la terminologie ont été mobilisées à travers divers travaux pluridisciplinaires. La partie suivante décrira un cadre méthodologique générique qui a été appliqué dans ces différents projets.

Indexation et classification automatique

- (i) Les données scientifiques sont par nature complexes et souvent spécialisées. C'est, par exemple, le cas des données qui s'intéressent au domaine agronomique et qui couvrent finalement un large spectre allant de l'étude biologique des plantes jusqu'aux facteurs environnementaux et sociétaux associés aux pratiques agricoles. L'indexation des données agronomiques est alors utile pour mieux appréhender et cartographier le patrimoine numérique scientifique disponible au sein des instituts de recherche en agriculture. Une telle tâche d'indexation se révèle cruciale pour une meilleure gestion des connaissances et favoriser l'ouverture des données agronomiques.
- (ii) Dans un tel contexte, les données textuelles issues des publications scientifiques recèlent des informations précieuses que des méthodes de fouille de textes peuvent mettre en lumière. Les processus mis en place sont souvent composés de deux phases. Dans un premier temps, ces méthodes consistent à extraire les descripteurs linguistiques les plus significatifs à partir de documents. La deuxième phase utilise ces termes pour, par exemple, classer automatiquement les documents

dans des catégories (« culture », « élevage », etc.). Cette classification repose sur le postulat suivant : si des documents possèdent de nombreux termes en commun alors ils peuvent être regroupés.

Dans Roche *et al.* (2015), les méthodes de fouille de textes ont été appliquées à un corpus issu d'Agritrop⁷ (articles, actes, ouvrages, chapitres d'ouvrages, thèses, HDR) datant de la période 2008-2012. Une analyse manuelle et qualitative avec des documentalistes et des experts en géographie et télédétection a montré que ces termes représentent des descripteurs thématiques utiles pour la recherche documentaire dans les masses de données aujourd'hui disponibles. Ils peuvent également être des indicateurs tout à fait pertinents pour la mise en correspondance de données hétérogènes (par exemple, enquêtes vs. publications scientifiques) dans le but, par exemple, de découvrir des connaissances nouvelles ou mettre en relation des chercheurs s'intéressant à des thématiques scientifiques proches (Roche *et al.*, 2015). Ce type de tâches d'indexation et de classification est également mis en œuvre dans le cadre du projet LEAP4FNSSA⁸ (2019-2022) qui intègre de nombreux documents (descriptions de projets, pages Web, publications, etc.). LEAP4FNSSA a pour ambition de créer et mettre au service de tous les acteurs concernés par la sécurité alimentaire et l'agriculture durable un outil KMS (*Knowledge Management System*) permettant d'optimiser les politiques, programmes, actions et mécanismes de financement (Martin *et al.*, 2021).

Veille automatique

- (i) Les activités de veille (sociétale, scientifique, technologique, économique, épidémiologique, etc.) consistent à collecter des informations stratégiques pour surveiller des situations et des problématiques. La veille est un élément crucial pour prévenir certains risques comme l'émergence de maladies connues ou inconnues.
- (ii) Dans le contexte de la « veille sanitaire internationale » de la plateforme nationale d'épidémiosurveillance en santé animale (plateforme ESA), le Cirad, l'Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES) et la direction générale de l'Alimentation (DGAl) développent, depuis 2013, un système de veille automatique du Web qui s'appuie sur une terminologie à différentes étapes du processus comme

⁶ <http://prosperologie.org/?sit=22>.

⁷ Base de données bibliographique du Centre de coopération internationale en recherche agronomique (Cirad).

⁸ Long-term EU-AU Research and Innovation Partnership for Food and Nutrition Security and Sustainable Agriculture.

cela sera détaillé en partie 4 de cet article. Le processus automatique de veille sanitaire en santé animale sur le Web (Valentin *et al.*, 2020) est décliné en 4 principales étapes : (1) la collecte automatique d'articles sanitaires via des requêtes Web (sur la base de mots-clés); (2) la classification automatique des articles collectés selon leur contenu par des méthodes d'apprentissage automatique : pertinents (articles qui décrivent des événements sanitaires liés à l'apparition des foyers de maladies exotiques ou nouvelles) et non pertinents (tout autre article); (3) l'extraction automatique de l'information sanitaire à partir des articles pertinents (maladie, date et lieu de l'évènement, signes cliniques, hôtes touchés, etc.); (4) l'analyse et l'évaluation du processus à l'aide de connaissances des experts du domaine. Les maladies actuellement « surveillées » par ce système de veille sont la peste porcine africaine, la grippe aviaire, la fièvre catarrhale ovine, la fièvre aphteuse, la maladie de Schmallenberg, etc. Le système est développé dans un cadre générique et permet la veille de nombreuses maladies comme la Covid-19 qui a été récemment intégrée dans un contexte *One Health* (Valentin *et al.*, 2021).

Mise en lumière de dires d'acteurs

(i) La notion de territoire, et plus spécifiquement d'aménagement du territoire, fait référence à différents concepts tels que les informations spatiales et temporelles, les acteurs, les opinions, l'histoire, la politique, etc. La caractérisation et la compréhension des perceptions d'un même territoire par les différents acteurs sont difficiles, mais particulièrement intéressantes dans une perspective d'aménagement du territoire et de politique publique territoriale. Cette problématique, abordée dans le contexte des territoires numériques, est très complexe et la mobilisation de méthodes de fouille de textes peut apporter des réponses intéressantes.

(ii) La notion de territoire repose en partie sur la présence dans les textes de descripteurs linguistiques associés à trois concepts, à savoir des entités thématiques, des entités spatiales et des sentiments. Dans ce contexte, des méthodes d'extraction de la terminologie peuvent être mises en place (*cf.* partie 4) pour des tâches mobilisant différents types de corpus :

- Identification automatique des opinions liées à l'aménagement du territoire à partir d'articles de presse du quotidien régional *Midi libre* relatifs au territoire de l'étang de Thau (Kergosien *et al.*, 2015).
- Identification du champ lexical sur la thématique de l'épandage aérien en production bananière et sur les

points de vue contrastés de deux ensembles d'acteurs – membres de la société civile *vs.* Union des groupements de producteurs de bananes de Guadeloupe et Martinique – (Bonin *et Roche*, 2019).

- Identification des informations liées à l'utilisation de techniques d'agriculture durable sur le territoire de Madagascar à partir d'un ensemble de données hétérogènes (thèses, mémoires, rapports, relevés, fiches techniques, etc.) (Fize *et al.*, 2019).

Dans une seconde étape, les différents éléments peuvent être mis en relation par des approches automatiques (Kergosien *et al.*, 2015 ; Fize *et al.*, 2019). Dans les travaux de Kergosien *et al.* (2015), nous avons identifié les couples candidats « opinion-entités nommées » (EN, entités spatiales et organisations). L'approche fondée sur des traitements statistiques se décompose en deux étapes principales : (1) indexation des documents, (2) identification des couples « opinion-EN ». Cela a permis de mettre en relief les couples tels que « magnifique/bassin de Thau », « belle/ville de Sète », etc. Cette mise en relation permet de mettre en avant que les opinions positives auraient plutôt tendance à être associées à des entités spatiales alors que les opinions plus négatives seraient davantage liées aux organisations. Ces tendances observées nécessitent néanmoins d'être étudiées de manière plus détaillée.

Construction de ressources terminologiques

Les ressources terminologiques/sémantiques à construire et/ou enrichir à partir de données textuelles peuvent être de deux ordres. Tout d'abord, comme dans les travaux que nous avons menés dans le contexte du projet GloFoodS Valorcarne⁹, ces données peuvent être très spécialisées (Roche *et al.*, 2017). Elles peuvent aussi être liées à des domaines assez généraux qui mobilisent des ressources sémantiques déjà existantes comme le décrit le paragraphe suivant.

Dans le cadre de la problématique de la mobilité qui possède un caractère éminemment pluridisciplinaire en sciences sociales mais également dans le domaine de la santé (par exemple, en épidémiologie) ou en agriculture (par exemple, pour les études liées à la mobilité animale), nous avons proposé une méthode intégrant différentes ressources sémantiques (en particulier, Agrovoc¹⁰) et des approches de fouille de textes pour amorcer la construction d'une terminologie (Roche *et al.*, 2018). Ce type d'approche a également été mis en place dans des

⁹ Optimiser les procédés de valorisation des produits carnés pour réduire les pertes et gaspillages.

¹⁰ <http://aims.fao.org/agrovoc>.

travaux liés à la construction de thésaurus d'agroécologie (Batifol *et al.*, 2020).

Dans l'ensemble de ces projets, une démarche générique a été instaurée pour mettre en avant la terminologie la plus significative associée à chaque tâche. Cette démarche qui est composée de 3 phases séquentielles est résumée dans la partie suivante.

Méthodologie

Construction d'un corpus

La collecte des données textuelles (corpus) peut être réalisée à travers le Web via des requêtes sur des moteurs de recherche (Bonin et Roche, 2019) ou par flux RSS (Valentin *et al.*, 2020). Ces corpus peuvent également être constitués à partir de ressources bibliographiques (Roche *et al.*, 2015 ; Roche *et al.*, 2018).

Dans ces différentes approches, l'utilisation des mots-clés est cruciale. Par exemple, pour la veille en épidémiologie animale, nous proposons une approche fondée sur des requêtes Web composées de termes tels que (i) des noms de maladies pour la veille propre à des maladies connues (*disease-specific*), (ii) des combinaisons entre des signes cliniques et des hôtes pour une veille syndromique (*non-specific*). Pour identifier les termes utiles pour la veille, nous avons effectué deux types d'actions qui se sont révélées parfaitement complémentaires : (i) nous avons recueilli, directement auprès des experts, les mots-clés qui caractérisent les maladies *via* des formulaires dédiés ; (ii) nous avons constitué un corpus et effectué une tâche de fouille de textes sur celui-ci permettant d'obtenir des mots-clés caractéristiques qui ont été validés par les experts (Arsevska *et al.*, 2016). Ces approches se sont révélées complémentaires et ont été réalisées *via* des approches Delphi qui sont décrites à la 3^e phase du processus. La place de l'expert pour cette phase de collecte des données est donc majeure.

Pour prendre en compte la discriminance des termes qui sera présentée dans le paragraphe suivant, les textes doivent être segmentés en documents. Cela peut être issu des segments structurels (documents, sections, etc.) ou thématiques.

Extraction de la terminologie

La deuxième phase de la démarche proposée consiste à identifier les termes pertinents pour les différentes tâches (*cf.* partie 3) par fouille de textes à partir de corpus dédiés. Dans un premier temps, les termes sont extraits des textes à l'aide des paramètres génériques propres au logiciel BioTex (Lossio-Ventura *et al.*, 2016) qui permet de

traiter des textes en anglais, français et espagnol. La motivation d'utiliser BioTex tient au fait que ce système permet de paramétrer librement l'extraction selon plusieurs critères numériques et méthodes décrits dans ce paragraphe. Le logiciel BioTex exploite à la fois des informations statistiques et linguistiques pour extraire une terminologie à partir de textes libres. Les termes candidats sont tout d'abord retenus s'ils respectent des patrons syntaxiques définis (adjectif-nom, nom-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un critère statistique est appliqué. Celui-ci mesure l'association entre les mots composant un terme en utilisant une mesure appelée *C-value* (Frantzi *et al.*, 2000) et en intégrant une pondération (*TF-IDF* ; *Term Frequency-Inverse Document Frequency*). Le but de *C-value* est d'améliorer l'extraction des termes composés alors que la pondération *TF-IDF* met en avant le pouvoir discriminant du terme candidat. L'extraction des termes peut être réalisée selon deux paramètres/fonctions :

- (i) Extraction (a) de termes composés, par exemple, *pastoral mobility*, *poultry trade network*, et (b) de termes mixtes (simples et composés), par exemple, *space*, *location*, *individual movement*, à partir des corpus.
- (ii) Sélection des *n* premiers termes classés par différentes mesures statistiques, en particulier les mesures *F-TFIDF-C* (Lossio-Ventura *et al.*, 2016) et *C-value* (Frantzi *et al.*, 2000) qui sont décrites ci-dessous.

F-TFIDF-C : Cette mesure, qui a récemment montré son bon comportement (Martin *et al.*, 2021), repose en partie sur le critère *TF-IDF* qui donne un poids plus important aux termes caractéristiques et discriminants d'un corpus (Salton et McGill, 1983). Pour attribuer un poids de *TF-IDF*, le produit entre la fréquence d'un terme (*Term Frequency*) et la fréquence inverse du document (*Inverse Document Frequency*) est réalisé. Ce dernier critère mesure le caractère discriminant du terme au regard de l'ensemble du corpus.

C-value : Cette mesure est fondée sur la fréquence des termes tout en privilégiant l'extraction de termes imbriqués et composés. Ce critère favorise les syntagmes (groupe de mots) n'apparaissant pas dans des termes plus longs. Par exemple, en veille en épidémiologie animale, le terme *african swine fever* est à privilégier au regard du syntagme *african swine*.

D'autres mesures peuvent également être mobilisées (Pazienza et Stellato, 2012) mais ces deux critères donnent des résultats à la fois pertinents et tout à fait complémentaires comme cela sera illustré dans le cas d'étude suivant. Dans le cadre d'une étude en épidémiologie animale, deux corpus en anglais ont été constitués (Roche et Arsevska, 2018) relatifs à la peste porcine africaine (*african swine fever*). Le premier

TEXT-MINING - TERMINOLOGY EXTRACTION

LEAP4FNSSA

Name	To complete...			
Institut				
Country				
Expertise/Position				
EVALUATION 1 (MACRO - EVALUATION) -> semantic for classification				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
To complete...				
Example(s) of terms (from All C-Value, All F_TFIDF_C, Multi C_Value, Multi All F_TFIDF_C lists)				
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
To complete...				
Rank	All C-Value	All F_TFIDF_C	Multi C_Value	Multi All F_TFIDF_C
1	food	water	food security	climate change
2	project	nutrition	burkina faso	fruit fly
3	africa	fruit	climate change	west africa
4	research	irrigation	fruit fly	sub-saharan africa
5	water	smallholder	nutrition security	market access
6	production	fish	west africa	small fish
7	development	oil	capacity building	sustainable intensification
8	and	food	food production	decision support
9	security	cassava	south africa	urban food

Fig. 1. Tableau d'évaluation des termes pour une tâche de classification de textes (projet LEAP4FNSSA).

corpus est composé de 69 textes en anglais décrivant des foyers épidémiologiques et 69 autres textes traitent d'aspects socioéconomiques liés à cette maladie. Les résultats¹¹ montrent que des termes très différents sont extraits selon les corpus et les mesures. Par exemple, le corpus lié aux conséquences économiques de la peste porcine africaine permet d'extraire un champ lexical associé aux institutions officielles (*European Commission, Russian Federation*), aux acteurs (*veterinary service, pig industry*), aux lieux à grande échelle (*West Africa*) et aux concepts importants (*animal health, food security*). Par ailleurs, les mesures fondées sur le critère de discriminance (IDF) qui extraient des termes pertinents pour le domaine (*animal health, swine flue*) ont tendance à mettre en avant des lieux à échelle plus réduite (*Lusaka province*) ou des acteurs plus précis (nom de personnes).

Validation

La validation des termes proposés par le système peut être réalisée à travers plusieurs approches. Dans un

premier temps, elle peut être effectuée en présentant des tableaux spécifiques aux experts (Fig. 1) ou des formulaires en ligne (Fig. 2 et 3). Elle peut également être réalisée dans le cadre d'ateliers dédiés qui permettent d'avoir un retour d'expérience des utilisateurs comme dans les travaux de Batifol *et al.* (2020). Cela a été mis en place dans le cadre du projet LEAP4FNSSA et le développement de la plateforme KEOPS (Martin *et al.*, 2021) pour lesquels des résultats de fouille de textes ont été proposés aux 9 participants et une discussion collective a été établie pour aboutir à des termes significatifs pour des tâches de classification et d'indexation (Roche *et al.*, 2020). Les consensus peuvent également être obtenus via des processus appelés Delphi comme cela a été mené par Arsevska *et al.* (2016). La méthode Delphi, par le biais d'une série d'actions et de rétroactions, permet d'obtenir des consensus pour déterminer la pertinence des termes relativement aux tâches dédiées. Par exemple, un terme peut être utile pour une tâche d'indexation d'un texte mais peu pertinent pour une tâche relative à l'analyse des dires d'acteurs. Par ailleurs, plusieurs niveaux de pertinence peuvent être identifiés. Afin de valider les termes extraits, plusieurs échelles de pertinence ont été identifiées et peuvent alors être prises en considération selon les projets.

¹¹ <http://agritrop.cirad.fr/594630/>.

	Very relevant	Relevant but to...	Not really relev...	Irrelevant	I don't know
urban agriculture	☐	☐	☐	☐	☐
food systems	☐	☐	☐	☐	☐
case studies	☐	☐	☐	☐	☐
agricultural res...	☐	☐	☐	☐	☐
wageningen uni...	☐	☐	☐	☐	☐
urban food	☐	☐	☐	☐	☐
small farms	☐	☐	☐	☐	☐
water harvesting	☐	☐	☐	☐	☐
sustainable agr...	☐	☐	☐	☐	☐
african countries	☐	☐	☐	☐	☐

Fig. 2. Formulaire en ligne utilisé pour valider des termes pour des tâches d'indexation dans le cadre du projet LEAP4FNSSA.

Fig. 3. Résultats de l'évaluation issus du formulaire (Fig. 2) dans le cadre du projet LEAP4FNSSA.

Tab 1. Investissement nécessaire de chaque type de compétence selon les différentes phases du processus.

	Compétences thématiques	Compétences méthodologiques
Constitution et normalisation du corpus	+ + +	+ +
Extraction de la terminologie	+	+ + + +
Évaluation des termes	+ + + +	+

Fig. 4. Cercle vertueux d'un processus de fouille de textes appliqué à des projets pluridisciplinaires.

- Le terme est pertinent pour la tâche : il peut alors être associé à des concepts importants (classes ou concepts).
- Le terme est pertinent mais très spécifique et pas nécessairement pertinent pour le domaine.
- Le terme est pertinent mais très général et pas nécessairement pertinent pour le domaine.
- Le terme est non pertinent.
- L'expert ne peut juger de la pertinence du terme.

Notons que la validation des termes lors de cette phase 3, pour des tâches de veille par exemple (cf. partie 3), est un élément crucial de la tâche de collecte à travers les mots-clés pertinents identifiés (phase 1). Cela permet alors d'affiner les tâches de collecte de la phase 1 et de respecter un cercle vertueux (Fig. 4).

Discussion et conclusion

Dans le cadre des différents travaux résumés en partie 3 et mis en œuvre à travers la démarche méthodologique présentée en partie 4, différentes compétences sont requises, en particulier des compétences thématiques liées aux cas d'étude et des compétences méthodologiques (fouille de textes) pour la mise en place et le paramétrage du processus. Pour

chacune des trois phases de la démarche (Fig. 4), les compétences requises sont plus ou moins importantes comme cela est proposé dans le tableau 1.

Il est cependant important de signaler que chaque étape nécessite la présence des deux compétences à des degrés différents. Il est essentiel que l'informaticien comprenne les problématiques thématiques et que l'expert des données et du domaine d'étude appréhende bien le principe des méthodes mobilisées pour choisir et discuter des résultats associés à chaque choix méthodologique (paramètres, mesures, etc.). Cela a été mis en place dans les différents projets et tâches associées qui sont présentés dans cet article.

Comme souligné dans l'état de l'art (cf. partie 2), les outils et plateformes de fouille de textes sont foisonnants dans la littérature. Certains plus complexes peuvent nécessiter des formations plus ou moins poussées ce qui peut constituer un frein dans leur utilisation. Pour certains projets pluridisciplinaires ou certaines études exploratoires et circonscrites, la mobilisation d'outils simples et flexibles peut être suffisante même si ces derniers peuvent avoir des limites intrinsèques.

Soulignons également que les méthodes d'intelligence artificielle sont aujourd'hui très efficaces et pertinentes pour des applications de fouille de textes (Piskorski *et al.*, 2020). Cependant, une limite reconnue de ces méthodes est l'absence d'explicabilité même si ces questions constituent aujourd'hui des axes de recherche prometteurs. Les approches, telles que celles présentées dans cet article, permettent néanmoins de mettre en relief les traces sémantiques véhiculées par les termes et constituent des axes d'étude complémentaires facilitant l'interprétabilité et l'analyse des résultats par les experts dans un contexte pluridisciplinaire.

Pour résumer, cet article a mis en avant de quelle manière l'entremêlement technique et scientifique, adossé à des questions de recherche thématiques, peut représenter une voie prometteuse des projets de recherche actuels et futurs. Dans ce contexte de dialogue pluridisciplinaire, les thématiciens peuvent mettre en relief des verrous scientifiques et méthodologiques que les experts en sciences des données pourront formaliser à l'aune de leur discipline.

Remerciements

Cette démarche a été menée à travers différents projets et expertises. Dans ce contexte, je remercie vivement les différents chercheurs qui ont contribué à consolider l'approche présentée dans cet article, en particulier Juan Antonio Lossio-Ventura, Clément Jonquet, Maguelonne Teisseire, pour le développement du logiciel BioTex, et les chercheurs, ingénieurs, étudiants qui l'ont appliqué dans les différents cas d'études décrits dans cet article : Elena Arsevska, Sarah Valentin (veille en épidémiologie animale), Muriel Bonin, Jacques Fize, Éric Kergosien (analyse des dires d'acteurs), Thierry Helmer, Pierre Martin (sécurité alimentaire).

Ces travaux ont été soutenus par la région Occitanie, par le Fonds européen de développement régional (FEDER) dans le cadre du projet SONGES (Science des données hétérogènes) et par l'Agence nationale de la recherche dans le cadre du programme Investissements d'avenir, référencé ANR-16-CONV-0004.

Références

- Arsevska E., Roche M., Hendrikx P., Chavernac D., Falala S., Lancelot R., Dufour B., 2016. Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web, *Computers and Electronics in Agriculture*, 123, 104-115, <https://doi.org/10.1016/j.compag.2016.02.010>.
- Aubin S., Hamon T., 2006. Improving term extraction with terminological resources, in Salakoski T., Ginter F., Pyysalo S., Pahikkala T. (Eds), *Advances in natural language processing, 5th international conference, FinTAL 2006 Turku, Finland, August 23-25, Proceedings*, Springer, 380-387.
- Barbier M., Cointet J.-P., 2012. Reconstruction of socio-semantic dynamics in sciences-society networks: methodology and epistemology of large textual corpora analysis. Communication au *Science and Democracy Network Annual Meeting*, 25-27 juin, Paris.
- Bartol T., 2009. Assessment of food and nutrition related descriptors in agricultural and biomedical thesauri, in Sartori F., Sicilia M.A., Manouselis N. (Eds), *Metadata and semantic research. Proceedings of the third international conference MTSR 2009, October 1-2, Milan, Italy*, Springer, 294-305.
- Batifol V., Aubin S., Hazard L., Theau J.-P., Bouffartigue C., Magrini M.-B., 2020. Thésaurus d'agroécologie : une approche par les usages et avis d'experts, *Cahiers Agricultures*, 29, 7, <https://doi.org/10.1051/cagri/2020004>.
- Bonin M., Roche M., 2019. Analyse des champs lexicaux des acteurs du territoire à partir de corpus textuels sur le Web : le cas des controverses autour de l'épandage aérien contre la cercosporiose du bananier en Guadeloupe, in Roche C. (Ed.), *TOTh 2019. Terminologie & ontologie : théories et applications*, Chambéry, Presses universitaires Savoie Mont Blanc, 293-307.
- Chateauraynaud F., 2003. *Prospéro. Une technologie littéraire pour les sciences humaines*, Paris, CNRS Éditions.
- Daille B., 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. Thèse de doctorat en sciences appliquées, Paris, Université Paris 7.
- David S., Plante P., 1990. De la nécessité d'une approche morphosyntaxique dans l'analyse de textes, *Intelligence artificielle et sciences cognitives au Québec*, 3, 140-154.
- Dobrov B., Loukachevitch N., 2011. Combining evidence for automatic extraction of terms, in Kuznetsov S.O., Mandal D.P., Kundu M.K., Pal S.K. (Eds), *Pattern recognition and machine intelligence. Proceedings of the 4th international conference PReMI 2011, June 27 – July 1, Moscow, Russia*, Springer, 235-241.
- Fize J., Roche M., Teisseire M., 2019. Mapping heterogeneous textual data: a multidimensional approach based on spatiality and theme, in El Yacoubi S., Bagnoli F., Pacini G. (Eds), *Internet science. 6th international conference, INSCI 2019, Perpignan, France, December 2-5, Proceedings*, Springer, 310-317, https://doi.org/10.1007/978-3-030-34770-3_25.
- Frantzi K.T., Ananiadou S., Mima H., 2000. Automatic recognition of multi-word terms: the C-value/NC-value method, *International Journal on Digital Libraries*, 3, 2, 115-130, <https://doi.org/10.1007/s007999900023>.
- Gauld C., Micoulaud-Franchi J.A., 2020. Analyse en réseau par fouille de données textuelles systématique du concept de psychiatrie personnalisée et de précision, *L'Encéphale*, <https://doi.org/10.1016/j.encep.2020.08.008>.
- Hazem A., Daille B., 2014. Semi-compositional method for synonym extraction of multi-word terms, in Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. (Eds), *Proceedings of the ninth international conference on language resources and evaluation (LREC 2014)*, European Language Resources Association, 2202-2207.
- Kennedy A., 2010. Automatically expanding the lexicon of Roget's thesaurus, in Farzindar A., Kešelj V. (Eds), *Advances in artificial intelligence. Canadian AI 2010. Lecture notes in computer science*, Springer, 410-413, https://doi.org/10.1007/978-3-642-13059-5_58.
- Kergosien É., Maurel P., Roche M., Teisseire M., 2015. SENTERRITOIRE pour la détection d'opinions liées à l'aménagement d'un territoire, *Revue internationale de géomatique*, 25, 1, 11-34, <https://doi.org/10.3166/RIG.25.11-34>.
- Kyriacopoulou T., Martineau C., Martinez C., 2018. UNITEX/GRAMLAB : plateforme libre basée sur des lexiques et des grammaires pour le traitement des corpus textuels, *Revue des nouvelles technologies de l'information*, vol. «Extraction et gestion des connaissances (EGC 2018)», 467-470.
- Laporte M.A., Mougenot I., Garnier E., 2012. ThesauForm-Traits: a web based collaborative tool to develop a thesaurus for plant functional diversity research, *Ecological informatics*, 11, 34-44, <https://doi.org/10.1016/j.ecoinf.2012.04.004>.
- Lossio-Ventura J.A., Jonquet C., Roche M., Teisseire M., 2016. Biomedical term extraction: overview and a new

- methodology, *Information Retrieval Journal*, 19, 1-2, 59-99, <https://doi.org/10.1007/s10791-015-9262-2>.
- Malanski P.D., Dedieu B., Schiavi S., 2021. Mapping the research domains on work in agriculture. A bibliometric review from Scopus database, *Journal of Rural Studies*, 81, 305-314, <https://doi.org/10.1016/j.jrurstud.2020.10.050>.
- Martin P., Helmer T., Rabatel J., Roche M., 2021. KEOPS: Knowledge ExtractOr Pipeline System, in Cherfi S., Perini A., Nurcan S. (Eds), *Research challenges in information science. 15th international conference RCIS 2021, Limassol, Cyprus, May 11-14, Proceedings*, Springer, 561-567.
- Nédellec C., Golik W., Aubin S., Bossy R., 2010. Building large lexicalized ontologies from text: a use case in automatic indexing of biotechnology patents, in Cimiano P., Pinto H.S. (Eds), *Knowledge engineering and management by the masses. EKAW 2010*, Springer, 514-523, https://doi.org/10.1007/978-3-642-16438-5_41.
- Névéal A., Grosjean J., Darmoni S.J., Zweigenbaum P., 2014. Language resources for French in the biomedical domain, in Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S. (Eds), *Proceedings of the ninth international conference on language resources and evaluation (LREC 2014)*, European Language Resources Association, 2146-2151.
- Paumier S., 2020. *Unitex 3.2. Manuel d'utilisation*, Université Paris-Est-Marne-la-Vallée, <https://unitexgramlab.org/reeses/3.2/man/Unitex-GramLab-3.2-usermanual-fr.pdf>.
- Pazienza M.T., Stellato A. (Eds), 2012. *Semi-automatic ontology development. Processes and resources*, Hershey, IGI Global.
- Piskorski J., Haneczok J., Jacquet G., 2020. New benchmark corpus and models for fine-grained event classification: to BERT or not to BERT?, in Scott D., Bel N., Zong C. (Eds), *Proceedings of the 28th international conference on computational linguistics, December 8-13, Barcelona*, International Committee on Computational Linguistics, 6663-6678.
- Roche M., 2020. How to define co-occurrence in a multidisciplinary context?, *Informatica*, 44, 3, 387-393, <https://doi.org/10.31449/inf.v44i3.2996>.
- Roche M., Arsevska E., 2018. *PADI-web: ASF corpora*, <https://doi.org/10.18167/DVN1/POIZMA>, CIRAD Dataverse.
- Roche M., Fortuno S., Lossio-Ventura J.A., Akli A., Belkebir S., Lounis T., Toure S., 2015. Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie, *Cahiers Agricultures*, 24, 5, 313-320, <https://doi.org/10.1684/agr.2015.0773>.
- Roche M., Teisseire M., Shrivastava G., 2017. *Valorcarn-TETIS: candidates for OTR (Ontological and Terminological Resource)*, <https://doi.org/10.18167/DVN1/KNFAGG>, CIRAD Dataverse.
- Roche M., Coste C., Falala S., Touré I., Bonnet P., 2018. Enrichissement et consolidation de ressources lexicales liées à la mobilité par des approches de fouille de textes, in Roche C. (Ed.), *TOTh 2018. Terminologie & ontologie : théories et applications*, Chambéry, Presses universitaires Savoie Mont Blanc.
- Roche M., Helmer T., Martin P., Chaminuka P., Dimitriou I., Csorba A., Lindsten A., Lundén T., van Boheemen P., 2020. *LEAP4FNSSA (WP3 – KMS): Terminology for KEOPS*, <https://doi.org/10.18167/DVN1/GQ8DPL>, CIRAD Dataverse.
- Salton G., McGill M.J., 1983. *Introduction to modern information retrieval*, New York, McGraw-Hill.
- Silberztein M., 2015. *La formalisation des langues. L'approche de NooJ*, Londres, ISTE Éditions.
- Smadja F., 1993. Retrieving collocations from text: Xtract, *Computational Linguistics*, 19, 1, 143-177.
- Turenne N., Barbier M., 2004. BELUGA : un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine. Première application au cas des maladies à prions, in Hébrail G., Lebart L., Petit J.-M. (Eds), *Extraction et gestion de connaissances (EGC 2004). Actes des quatrième journées « Extraction et gestion des connaissances »*, 20-23 janvier, Clermont-Ferrand, France, Toulouse, Éditions Cépaduès, 423-428.
- Vakkari P., 2010. How specific thesauri and a general thesaurus cover lay persons' vocabularies concerning health, nutrition and social services, in Gnoli C., Mazzocchi F. (Eds), *Paradigms and conceptual systems in knowledge organization. Proceedings of the eleventh international ISKO conference, 23-26 February, Rome, Italy*, Würzburg, Ergon Verlag, 299-307.
- Valentin S., Arsevska E., Falala S., De Goër J., Lancelot R., Mercier A., Rabatel J., Roche M., 2020. PADI-web: a multilingual event-based surveillance system for monitoring animal infectious diseases, *Computers and Electronics in Agriculture*, 169, 105163, <https://doi.org/10.1016/j.compag.2019.105163>.
- Valentin S., Mercier A., Lancelot R., Roche M., Arsevska E., 2021. Monitoring online media reports for early detection of unknown diseases: insight from a retrospective study of COVID-19 emergence, *Transboundary and Emerging Diseases*, 68, 981-986, <https://doi.org/10.1111/tbed.13738>.

Citation de l'article : Roche M. Comment mobiliser des approches de fouille de textes et d'extraction de la terminologie dans un contexte pluridisciplinaire ?. *Nat. Sci. Soc.* 29, 2, 223-232.