Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie

Mathieu Roche, Sophie Fortuno, Juan Antonio Lossio-Ventura, Amira Akli, Salim Belkebir, Thinhinan Lounis, Serigne Toure

Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie

Mathieu Roche[1,2]
Sophie Fortuno[1]
Juan Antonio Lossio-Ventura[2,3]
Amira Akli[3]
Salim Belkebir[3]
Thinhinan Lounis[3]
Serigne Toure[3]

[1] UMR TETIS (Cirad, Irstea, AgroParisTech)
Maison de la Télédétection
500, rue Jean-François Breton
34093 Montpellier Cedex 5
France
<mathieu.roche@cirad.fr>
<sophie.fortuno@cirad.fr>

[2] LIRMM (CNRS, Université de Montpellier)
860, rue de St Priest
34095 Montpellier Cedex 5
France
<juan.lossio@lirmm.fr>

[3] Université de Montpellier
Place Eugène Bataillon
34095 Montpellier Cedex 5
France

Résumé

Dans le contexte des masses de données textuelles liées à l'agriculture aujourd'hui disponibles, leur indexation devient un enjeu crucial pour les organismes de recherche. Une manière d'indexer au mieux les documents consiste à en extraire la terminologie. Cet article explore l'utilisation et la combinaison de méthodologies de fouille de textes afin de mettre en exergue, puis de publier dans des systèmes d'*open data*, les termes les plus adaptés issus de documents. Des expérimentations menées sur des données du CIRAD (Centre de coopération internationale en recherche agronomique pour le développement), montrent le bien-fondé de la démarche qui a permis d'extraire des termes à la fois nouveaux et pertinents.

Mots clés : documentation ; gestion des connaissances ; indexation d'information ; méthodes ; traitement des données.

Thèmes : méthodes et outils.

Abstract

Automatic extraction of keywords from scientific publications for indexing and *open data* in agronomy

With the large amounts of textual data related to agriculture now available, indexing becomes a crucial issue for research organizations. One way to index documents consists in extracting terminology. This paper investigates the use and combination of text mining methodologies to highlight and publish the most appropriate terms from documents in *open data* systems. Experiments conducted on CIRAD data, show the validity of the approach used to extract new and relevant terms.

Key words: data processing; documentation; indexing of information; knowledge management; methods.

Subjects: tools and methods.

Les données scientifiques sont par nature complexes et souvent spécialisées. C'est par exemple le cas des données qui concernent le domaine agronomique, qui couvrent 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

Tirés à part : M. Roche

doi: 10.1684/agr.2015.0773

Pour citer cet article : Roche M, Fortuno S, Lossio-Ventura JA, Akli A, Belkebir S, Lounis T, *et. al.*, 2015. Extraction automatique des mots-clés à partir de publications scientifiques pour l'indexation et l'ouverture des données en agronomie. *Cah Agric* 24 : 313-320. doi : 10.1684/agr.2015.0773

agriculture. Une telle tâche d'indexation se révèle cruciale pour mieux gérer les connaissances et favoriser l'ouverture des données agronomiques.

Dans un tel contexte, les données textuelles issues des publications scientifiques recèlent des informations précieuses que des méthodes de fouille de textes (FT, http://www.textmining.biz) peuvent mettre en exergue. Les processus de FT sont souvent composés de deux phases successives. Dans un premier temps, ces méthodes consistent à extraire les descripteurs linguistiques les plus significatifs à partir de documents. Les descripteurs linguistiques 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*. Ces derniers représentent le matériau de base afin d'associer une certaine sémantique aux documents. Par exemple, les termes « riz » et « irrigation » présents dans un document mettent en lumière une thématique liée à la « culture ». La deuxième phase du processus consiste à exploiter 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.

De telles problématiques de classification, d'indexation ou d'ouverture des données (*open data*) reposent donc, dans un premier temps, sur l'extraction des termes pertinents issus d'un ensemble de données textuelles appelé *corpus*. Dans le cadre de cette étude, nous nous intéressons à la mise en place d'une méthodologie afin d'extraire automatiquement les termes pertinents à partir des publications du CIRAD (Centre de coopération internationale en recherche agronomique pour le développement) et de l'unité mixte de recherche TETIS (Territoires, environnement, télédétection et information spatiale).

La section suivante décrit les méthodes d'extraction de la terminologie de la littérature et celles exploitées dans nos travaux. Celle d'après détaille une méthode de combinaison de différentes approches afin d'extraire des termes adaptés au domaine agronomique. Ensuite, est décrite une mesure originale de classement des termes qui ont été exploités pour des tâches propres à l'ouverture des données. Enfin, les résultats de nos approches sont analysés, puis mis en perspective.

Extraction automatique de la terminologie à partir de données textuelles

Cet article décrit une approche de FT afin d'extraire la terminologie. Dans un tel processus, il est nécessaire, au préalable, d'étiqueter les textes. L'étiquetage consiste à associer aux mots une fonction grammaticale (nom, verbe, etc.) en exploitant des informations lexicales et/ou contextuelles. Nos travaux s'appuient sur l'utilisation du Tree Tagger (Schmid, 1994). Un tel étiqueteur estime la probabilité qu'un mot ait une étiquette grammaticale en exploitant des arbres de décision binaires. Une fois l'étiquetage grammatical effectué, l'étape suivante consiste à extraire la terminologie.

État de l'art des méthodes d'extraction de la terminologie

Les termes constituent la base de ressources sémantiques, ou thésaurus, du domaine général (Kennedy, 2010 ; Vakkari, 2010) ou spécialisé, comme les sciences du vivant (Turenne et Barbier, 2004 ; Bartol, 2009 ; Névéol *et al.*, 2014). Leur construction peut être guidée :

– par consensus avec les experts (Laporte *et al.*, 2012) ;

– par les données nécessitant, par exemple, la mise en œuvre de méthodes de FT (Dobrov et Loukachevitch, 2011 ; Lossio Ventura *et al.*, 2014).

Notre travail utilise ce deuxième type d'approche. Les méthodes classiques d'extraction de la terminologie sont fondées sur des approches statistiques et/ou syntaxiques. 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 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). Contrairement à ACABIT, qui est essentiellement fondé sur des méthodes statistiques, LEXTER et SYNTEX s'appuient, en grande partie, sur une analyse syntaxique approfondie (Bourigault et Fabre, 2000). La méthode consiste à extraire les syntagmes nominaux maximaux. Ces derniers sont alors décomposés en termes de « têtes » et d'« expansions » à l'aide de règles grammaticales. Les termes sont alors proposés sous forme de réseau organisé en fonction de critères syntaxiques.

Afin de répondre à notre objectif qui consiste à effectuer des travaux préparatoires pour la réalisation d'un catalogue présentant et décrivant les bases de données et les publications associées à nos établissements de recherche, nous avons réalisé une extraction de la terminologie à travers deux applications (BioTex et GenTex) sur deux corpus (corpus CIRAD et TETIS). Le premier logiciel est dédié à l'extraction de termes spécialisés, en particulier en biomédecine (Névéol *et al.*, 2014). Le second s'intéresse à l'extraction de termes plus généraux, c'est-à-dire sans prendre en compte les spécificités syntaxiques des domaines de spécialité.

BioTex

BioTex est un logiciel qui exploite à la fois des informations *statistiques* et *linguistiques* pour extraire la

terminologie à partir de textes libres. Les informations statistiques apportent une pondération des termes candidats extraits. Cependant, la fréquence d'un terme n'est pas nécessairement un critère de sélection adapté. À titre d'exemple, le mot « agronomie » présent dans de très nombreuses publications du CIRAD se révèle très général et pas suffisamment discriminant. Ainsi, des mesures de discriminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes, peuvent être appliquées.

Mesure de discriminance

Pour effectuer une telle sélection, nos travaux s'appuient sur la pondération TF-IDF. Cette dernière donne un poids plus important aux termes caractéristiques d'un document (Salton et McGill, 1983). Ainsi, pour attribuer un poids de TF-IDF (*cf.* infra), il est nécessaire, dans un premier temps, de calculer la fréquence d'un terme (*term frequency*, TF). Ainsi, pour le document d_j et le terme t_i, la fréquence du terme dans le document est donnée par l'équation suivante :

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

où n_{ij} est le nombre d'occurrences du terme t_i dans d_j. Le dénominateur correspond au nombre d'occurrences de tous les termes dans le document d_j. La fréquence inverse de document (*inverse document frequency*, IDF) mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme et est définie de la manière suivante :

$$IDF_i = \log_2 \frac{|D|}{|d_j : t_i \in d_j|}$$

où $|D|$ représente le nombre total de documents dans le corpus et $|d_j : t_i \in d_j|$ représente le nombre de documents où le terme t_i apparaît. Enfin, la pondération finale s'obtient en multipliant les deux mesures :

$$TF\text{–}IDF_{ij} = TF_{ij} \times IDF_i$$

Mesures d'associations entre les mots

BioTex prend en compte deux facteurs pour extraire la terminologie. Tout d'abord, le logiciel extrait des termes selon des patrons syntaxiques définis (nom-adjectif, adjectif-nom, nom-préposition-nom, etc.). Après un tel filtrage linguistique, un autre filtrage statistique est appliqué. Celui-ci mesure l'association entre les mots composant un terme (par exemple, « agriculture familiale ») en utilisant une mesure appelée C-value (Frantzi *et al.*, 2000) tout en intégrant la pondération TF-IDF. Le but de C-value est d'améliorer l'extraction des termes complexes particulièrement adaptés pour les domaines de spécialité. Le critère mis en place permet de favoriser les termes n'apparaissant pas, de manière significative, dans des termes plus longs. Par exemple, dans un corpus spécialisé lié à l'ophtalmologie, Frantzi *et al.* (2000) montrent qu'un terme plus général comme « *soft contact* » est non pertinent alors que le terme plus long et donc plus spécifique de « *soft contact lens* » se révèle tout à fait pertinent.
Notons que le logiciel BioTex (version en ligne) propose deux types d'extraction (1 200 termes extraits au maximum) :
– termes composés uniquement ;
– termes simples et composés.
Ces termes sélectionnés sont ceux obtenant les meilleures pondérations statistiques selon les mesures précédemment présentées et détaillées dans Lossio Ventura *et al.* (2014).

GenTex

La seconde application expérimentée est la plateforme GenTex que nous avons mise en œuvre dans le cadre de nos travaux en FT. Elle permet d'extraire automatiquement des descripteurs linguistiques pertinents (mots respectant certaines caractéristiques) à partir d'un corpus préalablement étiqueté. Différentes fonctions et paramètres sont proposés :
– *fonction de pré-traitement* qui offre la possibilité de supprimer les « mots vides » (c'est-à-dire les mots fonctionnels, sans sémantique associée, comme les articles ou les prépositions) ;
– *fonction d'élagage* (*pruning*) qui permet à l'utilisateur de conserver les mots ayant un nombre d'occurrences supérieur à un seuil minimum fixé par l'utilisateur ;
– *fonction d'analyse linguistique* qui permet de présenter les mots associés à une étiquette grammaticale (par exemple, « cahiers/NOM ») ou sous forme lemmatisée (par exemple, « cahier ») ;
– *fonction de filtrage linguistique* qui offre la possibilité de sélectionner les mots selon leur catégorie grammaticale. Cela permet une analyse sémantique plus pointue d'un texte. Cette fonction permet de filtrer les sorties selon :
- toutes les catégories grammaticales se trouvant dans le corpus ;
- les adjectifs ;
- les noms communs et noms propres ;
- les verbes ;
- les adverbes ;

– *fonction de filtrage statistique* qui filtre les mots selon des mesures de qualité, par exemple, TF, IDF, DF, TF-IDF, TF-DF. Notons que la pondération DF correspond à l'inverse de l'IDF. Contrairement à BioTex, GenTex ne permet pas d'extraire des termes composés et *a fortiori*, ce système n'intègre pas de mesures d'association entre les mots (comme la mesure C-value).

Extraction de la terminologie du domaine agronomique par combinaison d'approches

Extraction terminologique et combinaison des approches BioTex et GenTex

Après exécution des méthodes précédemment décrites d'extraction de la terminologie selon des paramètres optimaux – discutés dans différents articles comme Lossio Ventura *et al.* (2014) – nous obtenons deux listes de termes (respectivement par GenTex et BioTex). Dans le but de comparer les résultats des deux types d'extraction,

nous avons étudié les descripteurs linguistiques communs des deux approches. De tels termes sont, *a priori*, les mots-clés les plus pertinents pour représenter nos documents. Une évaluation quantitative et qualitative des intersections et des termes spécifiques à chaque système sera détaillée en fin d'article.

Intégration d'un thesaurus de spécialité (Agrovoc)

Une fois les termes extraits et traités avec BioTex et GenTex, nous avons procédé à la validation des mots-clés à travers une ressource terminologique, Agrovoc (http://aims.fao.org/fr/agrovoc), thesaurus structuré et multilingue créé en 1980. Comme tout thésaurus, Agrovoc se structure selon des termes ou descripteurs qui comprennent un ou plusieurs mots représentant toujours un seul et même concept. Agrovoc est hiérarchisé selon 25 grands concepts tels que « activités », « processus », « méthodes », etc. Comparativement, le MeSH (*medical subject headings*) est structuré selon 16 catégories thématiques comme « anatomie », « organismes », « maladies », etc. Agrovoc concerne tous les domaines ayant un rapport avec l'agriculture, la foresterie, la pêche, l'alimentation et l'environnement. Il dispose d'une liste de plus de 32 000 éléments disponibles dans les six langues officielles de la FAO (Organisation des Nations Unies pour l'alimentation et l'agriculture) : anglais, arabe, chinois, espagnol, français et russe (information issue d'Agrovoc en date de mars 2015).

Combinaison des systèmes et des ressources

Afin de fiabiliser les résultats liés à l'extraction de la terminologie *via* BioTex et GenTex, nous avons effectué plusieurs intersections entre nos résultats et le thesaurus Agrovoc selon différentes stratégies décrites ci-dessous :
– (a) Combinaison des mots-clés simples (formés d'un seul mot) ;
• Intersection 1 : termes issus de l'intersection d'Agrovoc et BioTex ;
• Intersection 2 : termes issus de l'intersection de BioTex et GenTex ;
• Intersection 3 : termes issus de l'intersection de BioTex, GenTex et Agrovoc ;
– (b) Combinaison des mots-clés composés (formés de plusieurs mots) ;
• Intersection 4 : termes issus de l'intersection d'Agrovoc et BioTex ;
– (c) Combinaison des mots-clés simples et composés ;
• Intersection 5 : termes issus de l'intersection d'Agrovoc et BioTex.
Notons que les termes extraits avec nos deux systèmes et présents dans Agrovoc peuvent être considérés, de manière automatique, comme pertinents. Les résultats expérimentaux obtenus selon ces différentes stratégies sont détaillés plus loin.

Proposition d'une nouvelle mesure de pondération des termes agronomiques

Dans cette section, nous proposons une nouvelle mesure *LIDF-value* (*linguistic patterns, IDF, and C-value information*) qui consiste à pondérer puis classer les termes extraits selon les domaines de spécialité. Cette mesure, appliquée au terme t donnée ci-dessous, correspond au produit de trois pondérations : *IDF*, *C-value* (voir plus haut) et $P(t_{dom})$.

$$LIDF\text{-}value(t) = \mathrm{P}(t_{dom}) \times IDF(t) \times C\text{-}value(t)$$

L'intégration d'une pondération $\mathrm{P}(t_{dom})$ liée aux structures syntaxiques de chaque domaine *dom* (par exemple, la biomédecine ou l'agronomie) représente une originalité dans cette nouvelle fonction de rang *LIDF-value*. La pondération $\mathrm{P}(t_{dom})$ est apprise automatiquement en répertoriant les structures syntaxiques les plus significatives dans des thésaurus de spécialité. Par exemple, en biologie, la structure « NN CD NN NN NN » (où NN est un nom et CD un nombre) est assez significative ; elle est retrouvée dans 1107 cas et correspond à une probabilité de 0.27 à partir de la ressource UMLS (*unified medical language system*). Nous avons par ailleurs appris les patrons du domaine agronomique à partir des termes présents dans Agrovoc. Notre méthodologie est fondée sur les 200 patrons les plus significatifs pour l'anglais et le français. Notons que le domaine biomédical possède une forte présence de nombres dans les structures syntaxiques (correspondant à l'étiquette CD) contrairement au domaine agronomique. L'influence de cette nouvelle pondération propre au domaine agronomique à travers les expérimentations réalisées sur notre corpus du CIRAD est présentée plus loin.

Exploitation des ressources et publication de l'information

Travaillant en partenariat avec les pays du Sud, l'intérêt pour un organisme de recherche comme le CIRAD, est de disposer d'une représentation spatiale de son activité. Une piste d'exploration a été amorcée en combinant ces techniques d'extraction de la terminologie et la prise en compte d'informations spatiales extraites dans les articles. Les entités spatiales extraites dans les publications sont alors mises en relation avec des référentiels géographiques (par exemple, « GeoNames »). Notons que des approches spécifiques fondées sur des règles utilisant des indicateurs spatiaux (par exemple, « au sud de », « près de », etc.) peuvent être appliquées pour extraire les entités spatiales. Par ailleurs, des approches de désambiguïsation sémantique détaillées dans Kergosien *et al.* (2014) peuvent, au préalable, distinguer une entité spatiale d'une organisation (par exemple, selon le contexte, le mot « Brésil » présent dans un texte peut représenter une entité spatiale ou une entité politique).
L'ensemble du processus de FT que nous proposons permet de mettre en relief les articles scientifiques (représentés par un identifiant) traitant de sujets précis (termes extraits) en lien avec des entités spatiales (par exemple, un pays ou une ville). Les résultats peuvent alors être diffusés sur des plateformes dédiées à la publication de données en *open data* qui offrent certains modules de visualisation (*figure 1*).

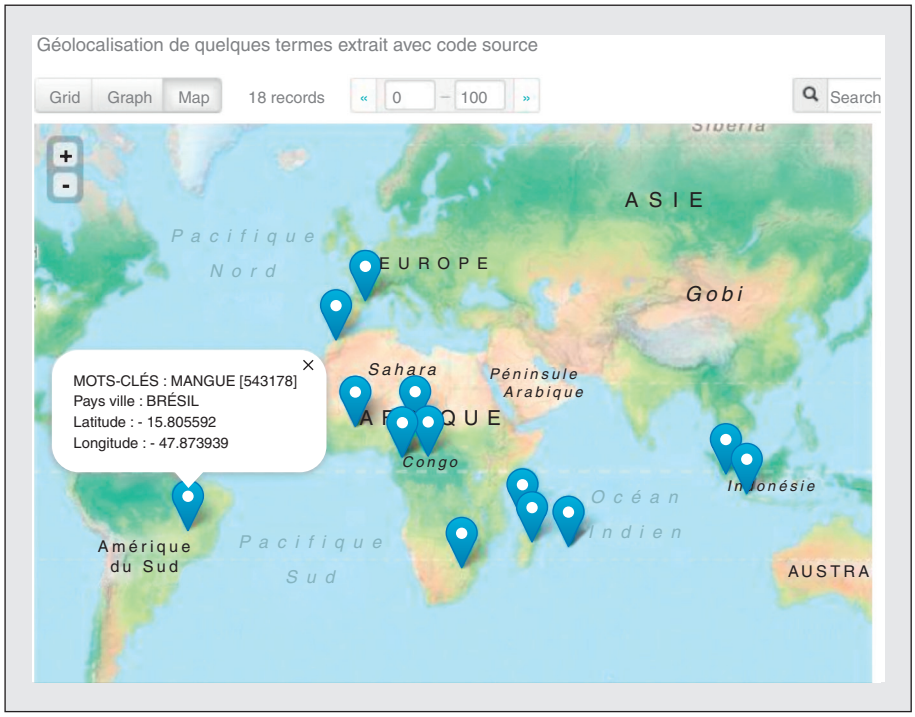

Figure 1. Visualisation spatialisée de mots-clés de publications scientifiques du CIRAD (ex. publication avec l'identifiant 543178) à travers la plateforme d'*open data* CKAN.

Figure 1. Spatial visualization of key words from scientific publications of CIRAD (for instance, publication with ID 543178) on CKAN *open data* platform.

Expérimentations

Protocole expérimental

Dans nos travaux, nous avons exploité des données issues d'Agritrop, l'archive ouverte des publications du CIRAD. Cette base de données documentaire est spécialisée en agronomie et filières de production animales et végétales des régions tropicales, subtropicales et méditerranéennes. Le corpus issu de nos expérimentations est une extraction effectuée sur les 250 000 références d'Agritrop, soit un sous-ensemble de résumés, articles, actes, ouvrages, chapitres d'ouvrages, thèses, habilitations à diriger des recherches, datant de la période 2008-2012 et écrits par des auteurs du CIRAD et de l'UMR TETIS. Le corpus en français du CIRAD a une taille de 4,26 Mo (2 145 documents), alors que le corpus de TETIS est 15 fois plus réduit (145 documents).
Le *tableau 1* présente un échantillon de 10 mots-clés simples et composés obtenus avec nos corpus et sur la base des différentes applications présentées plus haut.

Évaluation automatique des termes extraits

Le *tableau 2* résume les résultats obtenus avec chacune des intersections présentées précédemment. Ce tableau met en exergue plusieurs résultats intéressants selon le type d'intersections appliquées (mots simples, mots composés, mots simples et composés). L'analyse de ces résultats est détaillée ci-dessous.
Pour le corpus CIRAD, du croisement entre les termes simples extraits par BioTex et les termes simples d'Agrovoc, résultent 271 termes simples en commun, ce qui représente 28 % des mots-clés simples proposés par BioTex (intersection 1). La proportion est du même ordre avec le corpus TETIS. Par ailleurs, la grande majorité des mots-clés simples extraits par BioTex sont également extraits par GenTex (intersection 2). Le *tableau 2* met également en avant que 139 termes composés du corpus CIRAD (12 % des termes composés extraits avec BioTex) existent dans le thesaurus Agrovoc. Ces termes pourraient donc aider à l'enrichissement d'Agrovoc. En outre, pour le corpus TETIS, sur 1 200 termes composés extraits, seulement 17 existent dans Agrovoc. Cela peut s'expliquer par le fait que les textes issus du corpus TETIS sont plus spécifiques et/ou moins proches des thématiques que la ressource Agrovoc traite. Les expérimentations liées à l'intersection 5 mettent également ce point en avant.

Évaluation manuelle par des experts

En complément des travaux précédemment présentés, il s'agissait pour nous d'évaluer et indexer « en aveugle » les termes issus de ce processus de FT. Concrètement, nous avons effectué une analyse rigoureuse des méthodes d'indexation automatique

Tableau 1. Échantillon de mots-clés (en français) avec différents systèmes d'extraction de la terminologie.

Table 1. Sample of keywords (in French) with different terminology extraction systems.

Termes issus de BioTex (termes composés)	Termes issus de GenTex
systèmes de culture	développement
développement durable	gestion
canne à sucre	étude
systèmes de production	culture
sécurité alimentaire	marché

Tableau 2. Nombre de termes obtenus avec les différentes stratégies d'intersection.

Table 2. Number of terms obtained with different intersection strategies.

Type d'intersection	Corpus CIRAD	Corpus TETIS
Intersection 1	271 (28% BioTex simples)	118 (25% BioTex simples)
Intersection 2	894 (92% BioTex simples et GenTex)	418 (87% BioTex simples et GenTex)
Intersection 3	261 (29% d'intersection 2)	105 (25% d'intersection 2)
Intersection 4	139 (12% de BioTex composés)	17 (1% de BioTex composés)
Intersection 5	321 (27% de BioTex simples et composés)	135 (11% de BioTex simples et composés)

via l'évaluation manuelle de 140 termes-candidats issus du processus de FT par une experte documentaliste. Pour ce faire, nous avons introduit la notion de pertinence selon différents critères (sélectivité, spécificité, objectivité, cohérence, compatibilité structurelle, indépendance contextuelle). Le protocole méthodologique a abouti à la validation manuelle des termes-candidats selon ces critères. Cette évaluation manuelle a mis en avant que 82 % (resp. 76 %) des mots-clés simples issus du corpus CIRAD (resp. TETIS) sont pertinents en termes de sélectivité et spécificité. Ce taux de pertinence est à plus de 93 % pour les mots-clés composés extraits des deux corpus. La sélectivité correspond au fait que le terme représente au mieux le contenu du document. La spécificité correspond au niveau de détail et de précision des termes. Globalement, les résultats montrent une plus grande richesse lexicale des publications CIRAD.

Pour mener une étude avec une expertise thématique de notre approche, deux chercheurs du CIRAD qui sont membres de l'UMR TETIS ont été sollicités. Des résumés des articles en français (*encadré 1*) de ces chercheurs en géographie et télédétection ont constitué nos corpus (entre 15 et 20 résumés pour chaque expert). Les 100 premiers termes-candidats composés et classés avec la mesure *LIDF-value* ont alors été présentés aux chercheurs. L'évaluation a montré des résultats très différents, puisque dans le domaine de la géographie, 93 % des termes composés candidats ont été jugés pertinents, alors qu'en télédétection, seuls 56 % des termes ont été validés au niveau disciplinaire. Cette différence pourrait s'expliquer par le fait que les mots-clés très généraux (par exemple, « composante technologique », « informations pertinentes ») ou très spécifiques (par exemple, « nuisances environnementales », « inventaires forestiers ») ont été jugés non pertinents par l'expert en télédétection, contrairement au domaine de la géographie où les

Encadré 1

L'exemple ci-dessous met en exergue deux résumés d'articles de deux chercheurs dans le domaine de la géographie et de la télédétection. Les 15 meilleurs termes-candidats extraits sur la base de *LIDF-value* sont soulignés. Les termes sélectionnés qui sont également présents dans Agrovoc sont en gras (identification réalisée au cours du premier semestre 2015). Sur les 15 candidats, seuls 3 ont été jugés non pertinents par les experts (« décision », « enjeu fort », « décision individuelle »). D'autres termes intéressants sont extraits en considérant un seuil plus élevé de termes-candidats à retenir (« politiques agricoles », « mesures agri-environnementales », « production agricole », etc.).

• Résumé 1 (télédetection)

L'information spatialisée tient une place déterminante dans toutes les prises de **décision** liées à la **production** agricole et à l'environnement, depuis la prise de décision individuelle d'un agriculteur sur son exploitation jusqu'à la définition de stratégies globales des filières agricoles ou des régions. L'acquisition d'images par capteurs embarqués satellitaires ou sur systèmes légers (ULM, drone) et leur interprétation « agronomique » sous forme de cartes d'indicateurs pertinents pour les professionnels de la filière constituent un enjeu fort pour l'aide à la gestion de la **production** cannière.

• Résumé 2 (géographie)

La notion de service environnemental (SE) permet d'analyser la prise en compte de la question environnementale dans les politiques publiques. Nous avons conduit cet exercice dans deux territoires insulaires tropicaux soumis à la législation européenne, la Réunion et la Guadeloupe. Nous avons décrit le paysage général des instruments de politiques agricoles qui visent à protéger l'environnement, puis nous avons reconstruit la trajectoire de l'un de ces dispositifs, les mesures agri-environnementales. Le bilan de l'**appropriation** des dispositifs et de leur nouvelle philosophie reste mitigé et paradoxal. Afin de répondre facilement aux exigences administratives européennes, les institutions réunionnaises et guadeloupéennes continuent de promouvoir l'intensification de ces agricultures insulaires. Nous discutons enfin de l'intérêt d'une **appropriation** du concept de SE pour conduire davantage ces **agricultures** dans la voie du **développement durable**.

Tableau 3. Précision avec la mesure *LIDF-value* calculée à partir des corpus anglais et français.

Table 3. Precision with LIDF-value computed from English and French corpora.

Corpus	Anglais		Français	
Patrons	**Biomédicaux**	**Agronomiques**	**Biomédicaux**	**Agronomiques**
P@100	88,0 %	**92,0 %**	51,0 %	**56,0 %**
P@200	85,5 %	**87,5 %**	51,0 %	**53,5 %**
P@1000	68,2 %	**76,6 %**	33,6 %	**36,7 %**
P@5000	32,0 %	**34,0 %**	13,0 %	**13,5 %**

LIDF = *linguistic patterns - inverse document frequency*

candidats spécifiques et généraux ont été évalués positivement. Notons que la pertinence des termes extraits est souvent liée aux tâches à réaliser (indexation, recherche documentaire, extraction d'information, enrichissement de thésaurus), qui doivent être rigoureusement décrites aux experts dans un processus d'extraction et de validation terminologique.

Évaluation de la nouvelle mesure de classement des termes agronomiques

Afin d'évaluer notre nouvelle fonction de rang présentée plus haut, nous avons extrait la terminologie à partir d'un sous-ensemble du corpus CIRAD en anglais (156 résumés) et en français (84 résumés), tout en prenant en compte des patrons biomédicaux et agronomiques. Le *tableau 3* présente une comparaison de l'extraction des mots-clés, extraits à partir de ces corpus, en appliquant les pondérations $P(t_{dom})$ aux relations syntaxiques selon les domaines de spécialité (biomédecine *versus* agronomie). Nous avons calculé la pertinence des premiers termes extraits avec *LIDF-value* en calculant la précision (P) selon les n premiers termes (simples et composés) classés avec *LIDF-value*, noté P@n. La précision calcule la proportion de termes pertinents retournés par notre système. Dans nos expérimentations, nous considérons les termes comme pertinents s'ils sont présents dans Agrovoc. Nous pouvons relever que les résultats qui s'appuient sur des structures syntaxiques du domaine agronomique donnent de meilleurs résultats pour l'anglais et le français.

Conclusion et perspectives

Pour traiter les masses de données aujourd'hui disponibles (« l'infobésité »), la problématique de recherche du *big data* est classiquement mise en avant avec les 3V qui la caractérisent : volume, variété et vélocité. Toutes ces problématiques ouvrent de nouvelles disciplines de recherche comme la *science des données* qui mêle, entre autres, informatique, mathématiques, statistiques, visualisation et fouille de données/textes.

Les méthodes de FT décrites dans cet article ont été appliquées et adaptées à l'aide de divers pré- et post-traitements pour explorer le domaine agronomique. Une analyse manuelle et qualitative avec des documentalistes et des experts en géographie et télédétection a montré que, globalement, les termes issus des processus de FT se sont révélés appropriés. 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 *versus* publications scientifiques) dans le but, par exemple, de découvrir des connaissances nouvelles ou de mettre en relation des chercheurs s'intéressant à des thématiques scientifiques proches. ■

Remerciements

Les auteurs remercient la DIST (Délégation à l'information scientifique et technique) du CIRAD pour l'accès aux corpus, ainsi que Xavier Augusseau, Agnès Bégué et Odile Aptel-Barral qui ont participé à la phase d'évaluation. Ce travail est partiellement financé par le projet SIFR (ANR-12-JS02-0010) et par le programme de bourses FINCyT du Pérou.

Références

Bartol T, 2009. Assessment of food and nutrition related descriptors in agricultural and biomedical thesauri. In: *Metadata and semantic research.* Serie communications in computer and information science. Berlin : Springer Berlin Heidelberg : 294-305.

Bourigault D, Fabre C, 2000. Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires* 25:131-51.

Daille B, 1994. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques.* Thèse de doctorat. Université Paris 7. http://www.bdaille.com/index.php?option=com_docman&task=doc_download&gid=8&Itemid=

David S, Plante P, 1990. De la nécessité d'une approche morpho syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec* 3:140-54.

Dobrov B, Loukachevitch N, 2011. Combining evidence for automatic extraction of terms. In: *Pattern recognition and machine intelligence.* Serie Lecture notes in computer science. Berlin : Springer Berlin Heidelberg : 235-41.

Frantzi K, Ananiadou S, Mima H, 2000. Automatic recognition of multi-word terms: the C-value/NC-value method. *Inernational Journal on Digital Libraries* 3:115-30.

Hazem A, Daille B, 2014. Semi-compositional method for synonym extraction of multi-word terms. *Language Resources and Evaluation*; 1202-7.

Kennedy A, 2010. Automatically expanding the lexicon of Roget's thesaurus. In: *Advances in Artificial Intelligence.* Serie Lecture notes in computer science. Berlin : Springer Berlin Heidelberg : 410-3.

Kergosien E, Laval B, Roche M, Teisseire M, 2014. Are opinions expressed in land-use planning documents? *International Journal of Geographical Information Science* 28:739-62.

Laporte MA, Mougenot I, Garnier E, 2012. ThesauForm-Traits: a web based collaborative tool to develop a thesaurus for plant functional diversity

research. *Ecological informatics (Special Issue)* 11:34-44.

Lossio Ventura JA, Jonquet C, Roche M, Teisseire M, 2014. Towards a mixed approach to extract biomedical terms from text corpus. *International Journal of Knowledge Discovery in Bioinformatics* 4:1-15.

Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P, 2014. Language resources for french in the biomedical domain. *Language Resources and Evaluation*; 2146-51.

Salton G, McGill MJ, 1983. *Introduction to modern information retrieval*. McGraw-Hill College, 1983.

Schmid H, 1994. Probabilistic part-of-speech tagging using decision trees. *New Methods in Language Processing*; 44-9.

Smadja F, 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19:143-77.

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. Extraction et Gestion de Connaissances. *Revue des nouvelles technologies de l'information*: 423-8.

Vakkari P, 2010. How specific thesauri and a general thesaurus cover lay persons' vocabularies concerning health, nutrition and social services. In: *Paradigms and conceptual systems in knowledge organization: Proceedings of the Eleventh International ISKO Conference 23-26 February 2010 Rome, Italy*. Ergon Verlag : 299-307.