Cahiers
#Costech

> Souha Kefi
Oumayma Sboui
> Michel J.-F. Dubois
> Davide Rizzo

Le bricolage, au cœur des transitions agricoles vers la durabilité des agroéquipements

Les points critiques d'une fouille de données de vidéos YouTube

> #Numéro 6
> Evolution agrotechnique contemporaine
> Working papers
> Agriculture et technologie - > Développement durable

Citer cet article

Souha Kefi., Oumayma Sboui., Dubois, Michel J.-F.., Davide Rizzo. "Le bricolage, au cœur des transitions agricoles vers la durabilité des agroéquipements. Les points critiques d'une fouille de données de vidéos YouTube", 4 mai 2023, *Cahiers Costech*, numéro 6.
DOI https://doi.org/10.34746/cahierscostech169 -
URL https://www.costech.utc.fr/CahiersCostech/spip.php?article169

Résumé

Les progrès de l'agromachinisme devraient améliorer les conditions de travail des agriculteurs et répondre aux demandes sociétales pour une utilisation plus efficace et plus saine des ressources environnementales. Cependant, ces progrès peuvent réduire la maîtrise des équipements par les agriculteurs. Les agriculteurs peuvent adapter leurs équipements par le biais du « bricolage », qui est un ensemble de petites modifications non structurelles des équipements pour répondre aux besoins culturaux. Les informations sur les bricolages sont difficiles à collecter, mais les médias sociaux, en particulier YouTube, peuvent être une source utile sur ce sujet. Cet article présente une étude exploratoire qu'a été menée en deux parties : créer une base de données des créateurs de contenu agricole sur les médias sociaux, puis pour ensuite récupérer et analyser les vidéos YouTube liées au bricolage, à l'équipement agricole et aux principales opérations agricoles. La fouille de données a porté à la fois sur les transcriptions disponibles ou automatiques des vidéos, ainsi que sur les métadonnées associées à chaque vidéo. Les résultats obtenus montrent que les vidéos publiées en rapport avec le bricolage concernent les pratiques en rapport avec le sol, ce qui suggère fortement une approche agroécologique. La nouveauté de cette étude reste en premier lieu l'originalité de la méthodologie employée.

Plan

1 - Introduction

Le verbe « bricoler » indique un ensemble d'actions fragmentées qui réclament une ingéniosité et de l'habileté manuelle dans l'exploitation de moyens éclectiques (La langue française 2022). Le « bricolage », qui est le terme pour définir l'ensemble de ces actions, se pose ainsi en contraste avec un travail ordonné réalisé avec des moyens prédéfinis qui caractérise l'œuvre des ingénieurs (Rogers 2015).

Sur le plan conceptuel, plusieurs auteurs ont défini le bricolage (Odin and

Thuderoz 2010 ; Bendall 1983 ; Mélice 2009 ; Kincheloe 2011), jusqu'à en faire un cadre pour caractériser certaines approches de la recherche qualitative (Denzin and Lincoln 2017). Des exemples d'adoption du concept de bricolage concernent la médecine (M'zoughi 2021), l'enseignement (Hatton 1989), l'art (Stinchfield, Nelson, and Wood 2013), la psychologie (Trouvé 2020) ou encore l'entreprenariat (Senyard, Baker, and Davidsson 2009 ; Archer, Baker, and Mauer 2009).

L'étude que nous présentons porte sur le bricolage des agroéquipements par les agriculteurs. Nous allons nous rattacher à la définition de bricolage développée en anthropologie afin d'éclairer le sens et la pertinence du concept de bricolage pour notre étude. Lévi-Strauss (1962) aborde le concept de bricolage en ces termes : « […] une forme d'activité subsiste parmi nous qui, sur le plan technique, permet assez bien de concevoir ce que, sur le plan de la spéculation, put être une science que nous préférons appeler « première » plutôt que primitive, : c'est celle communément désignée par le terme de bricolage ». Plus particulièrement, Lévi-Strauss distingue deux types de bricolage :
– 1) Le bricolage technique « associé au travail des mains et associant des moyens détournés par rapport à ceux d'un artisan » (Lévi-Strauss 1962). Par l'étude de l'étymologie du terme français 'bricoler', Lévi-Strauss souligne la référence à un mouvement rapide qui impliquerait l'utilisation de moyens d'immédiate disposition pour atteindre les objectifs préfixés. Quatre profils de bricoleurs peuvent être identifiés par le rapport que bricoleur a aux projets et par la façon dont il travaille avec ce qui est à portée de main pour répondre aux événements de manière ad hoc. (a) L'optimisateur, qui se demande comment les matériaux à portée de main pourraient être utilisés, en pliant le degré d'achèvement du projet sur la réserve fixe des matériaux sans motivation à en acquérir de nouveaux. (b) Le collectionneur, pour lequel les moyens sont collectés non pas en vue d'une utilisation particulière, mais dans l'espoir qu'ils puissent s'avérer utiles. (c) L'expérimenté, dont les moyens sont déterminés sur la base d'expériences passées ; ses matériaux sont finis et chaque élément est ouvert à une variété d'usages et permet au bricoleur de fonctionner sans " l'équipement et le savoir de tous les métiers et professions" (Lévi-Strauss 1962). (d) L'opportuniste, qui se limite à une réorganisation ou une improvisation, incluant de nouvelles utilisations, de l'ensemble des moyens existants pour créer de nouvelles structures en réponse ad hoc à l'environnement. De fait, dans cette étude, nous nous intéresserons aux profils (c) et (d) qui correspondent aux agriculteurs bricoleurs.

– 2) Le bricolage conceptuel, qui concerne la science du concret. Ce mode de pensée antérieur à la science de l'abstrait car portant sur l'explication de catégories concrètes accessibles par des perceptions sensorielles (outils, espèces, nourriture, ...) est une pratique analogue à celle du bricoleur sur le plan technique. En revanche, la méthode du praticien qui utilise les sciences de l'abstrait est analogue à celle de l'ingénieur. Le bricolage aurait donc un aspect empirique mais sous-tendu et soutenu par le but d'atteindre un objectif.

Le bricolage semble être un concept multidimensionnel et ses manifestations très nombreuses sont difficiles à capter de manière représentative. Dans sa dimension technique, nous nous intéresserons au bricolage en agroéquipements comme manifestation d'un besoin d'adaptation des moyens utilisés par les agriculteurs pour réaliser des actions techniques spécifiques propres à l'agroécologie. Ce besoin concerne tant des adaptations agronomiques, que climatiques, mécaniques ou économiques, en lien implicite ou explicite avec les demandes de transition vers la durabilité. En effet, face à une situation inédite, la posture du bricoleur peut être en meilleure adéquation que la posture de l'ingénieur grâce à l'exploitation des moyens existants reconfigurés ou ajoutés pour une réponse ad hoc à l'environnement.

Afin de tracer un premier portrait des bricolages des agroéquipements nous avons choisi d'explorer les témoignages vidéo des agriculteurs sur YouTube. Ce réseau social permet de poster et visionner librement du contenu audiovisuel de durée très variable. Il est ainsi devenu un moyen efficace pour partager ou apprendre des pratiques, tel que le bricolage, qui ne sont pas standardisées, avec un accès sur-mesure à l'information (Casey et al. 2022). De plus, la simplification de production de contenus audiovisuels, notamment par l'ample disponibilité à des smartphones dotés de caméra et microphone, a permis à une nouvelle génération d'agriculteurs passionnés connectés (les agri-youtubeurs) d'atteindre un réel succès sur les réseaux sociaux (Bentley et al. 2019 ; Rénier 2022 ; Rénier et al. 2022). Nous ciblerons donc des agri-youtubeurs qui parlent face à la caméra de leurs bricolages d'agroéquipements. Nous ne ciblerons pas spécifiquement des bricoleurs orientés en agroécologie dans un contexte de transition. Nous supposons que cette recherche nous permettra de l'obtenir comme résultat.

Nous insisterons d'abord sur la partie 2, Matériel et méthodes, car les nouveaux outils employés ici font partie de l'étude et les résultats

obtenus comprendront la validation de cette approche expérimentale nouvelle. La partie 3, Résultats, comprendra donc aussi ceux qui valident ou infirment cette nouvelle exploration. Puis nous présenterons les résultats obtenus par cette méthode. Dans la partie 4, nous discuterons de l'intérêt et des limites de cette nouvelle approche. Finalement, nous conclurons dans la partie 5, Conclusions.

2 - Matériel et méthodes

La fouille de données (data mining en anglais) est définie comme la pratique consistant à examiner une base de données afin d'en générer de nouvelles informations. Cette pratique vise à mettre en exergue d'éventuelles corrélations significatives entre les données observées (Zaki and Wong 2004 ; Osman 2019 ; Obenshain 2004 ; Jun Lee and Siau 2001 ; Pujari 2001). Dans les paragraphes suivant nous détaillons les quatre grandes étapes qui composent l'analyse par fouille de données : collecte, préparation, analyse et modélisation (Zaki and Wong 2004 ; Osman 2019 ; Jun Lee and Siau 2001 ; Pujari 2001). Pour l'exécution de chacune de ces étapes, nous avons utilisé des ensembles de fonctions pré-écrites, ce qui nous évite de réécrire de zéro ces fonctions. En informatique des ensembles cohérents de fonctions sont nommées bibliothèques ; chaque bibliothèque est caractérisée par des cas d'usage et pour des environnements d'application spécifiques, écrits dans le langage de programmation adapté. Toutes les bibliothèques que nous avons utilisées étaient écrites dans le langage de programmation python. Nous garderons les termes techniques anglais dans les cas d'absence de mot français correspondant.

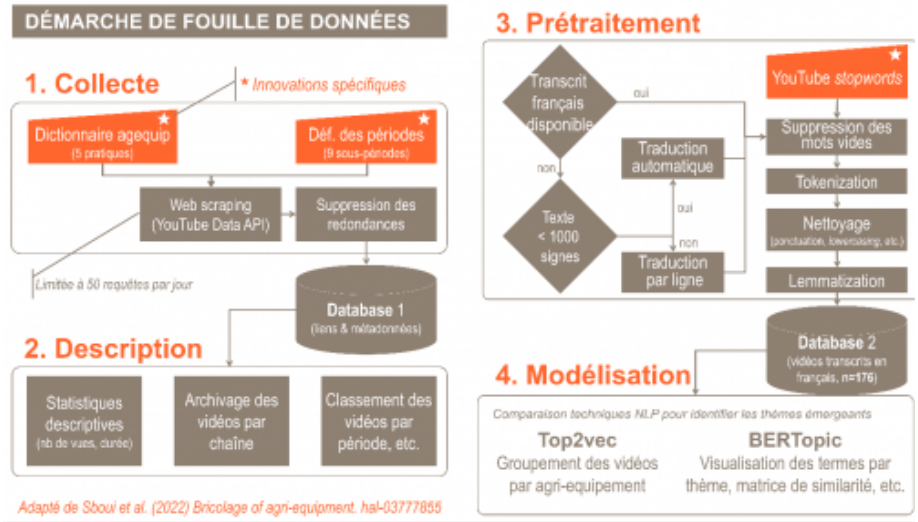

Figure 1. Processus d'extraction des transcriptions, de traduction et de traitement du texte. Adapté de Sboui et al. (2022).

Figure 1. Processus d'extraction des transcriptions, de traduction et de traitement du texte. Adapté de Sboui et al. (2022).

Figure 1. Processus d'extraction des transcriptions, de traduction et de traitement du texte. Adapté de Sboui et al. (2022).

2.1 - Collecte et description des données

Le matériel brut pour notre analyse se compose de deux parties : les vidéos et les données sur les vidéos, dites « métadonnées ». Les métadonnées des pages YouTube sont les informations textuelles descriptives qui définissent une vidéo, elles indiquent aux internautes tous les renseignements d'une vidéo (Baudet and Point 2017 ; Rangaswamy et al. 2016 ; Aggarwal, Agrawal, and Sureka 2014 ; Palod et al. 2019 ; Agarwal et al. 2017 ; Marathe and Shirsat 2015). La collecte a donc visé trois grandes étapes : d'abord, l'identification des vidéos pertinente par le web scraping, ensuite l'extraction des métadonnées et, enfin, l'extraction des transcriptions des vidéos.

Le web scraping prend son nom du verbe anglais to scrape, qui signifie « gratter » avec référence à l'extraction ciblée du contenu du web de façon structurée afin de le réutiliser pour des intérêts spécifiques (Glez-Peña et al. 2014 ; Khder 2021 ; Singrodia, Mitra, and Paul 2019 ; Karthikeyan et al. 2019 ; Zhao 2017). Le scraping (ou extraction) peut être réalisé via un programme, un logiciel automatique ou un autre site ; il est également appelé crawling quand il est assuré par des bots ou robots, qui répètent périodiquement l'extraction à l'instar, par exemple, des moteurs de recherche tels que Google (Glez-Peña et al. 2014 ; Khder 2021 ; Singrodia, Mitra, and Paul 2019).

Nous nous sommes concentrés sur trois des principales bibliothèques python existantes (Khder 2021 ; Lawson 2015 ; Mehta et al. 2020) pour réaliser du web scraping de liens :
– Selenium, adaptée à tout type de page web et permettant des commandes en plusieurs langages (C, Haskell, Java, JavaScript, Objective-C, Perl, PHP, Python, R et Ruby) tout en automatisant les interactions.
– Beautiful Soup, adaptée à tout type de page web, permettant la recherche d'éléments spécifiques grâce à l'analyse syntaxique de documents HTML et XML

– Pytube, la plus adaptée à notre étude car spécialisée seulement sur les pages YouTube, légère et sans dépendance, permettant de télécharger des vidéos à partir du web et d'autres fonctionnalités spécifiques.

Pour assurer une bonne extraction des liens reliés à la thématique, un dictionnaire a été construit. Ce dictionnaire identifie les pratiques des grandes cultures et les agroéquipements associés (Tableau 1). L'élaboration du tableau se base sur trois sources : 1) différents échanges avec des agriculteurs, 2) échanges avec les membres de notre équipe de recherche, 3) littérature grise relative aux agroéquipements.

Pratiques	Agroéquipements
Préparation du sol	Cultivateurs (dents, disques) Herses et fraises rotatives Déchaumeurs (dents, disques) Décompacteurs Charrues à socs (simple, réversible (principe « Brabant »), portée, semi-portée) Pulvérisateurs
Semis	Semoirs portés Semoirs tractés (trémie simple ou double) Semoirs combinés (préparation du sol et semis) Semoirs à trémies frontales Semoirs de précision Semoirs de semis direct
Désherbage	Outils de désherbage mécanique Herse étrille Houe rotative Bineuse (à socs, à étoiles, à doigts)
Récolte	Moissonneuse batteuse Faucheuses (À assiettes, à tambour)
Ensilage	Faneuse Andaineur Presse

Tableau 1. Dictionnaire des pratiques agricoles et les agroéquipements associés

L'accès aux vidéo YouTube a été réalisé au travers d'une interface de programmation d'application, ou "Application Programming Interface" en anglais (API). L'API est un ensemble de protocoles et de fonctions

permettant la communication entre logiciels et/ou bases de données. Les API servent à faciliter l'intégration et la communication entre différents systèmes, avec un contrôle très fin des droits d'accès. Dans notre cas d'étude, l'API YouTube, extrait via la plateforme YouTube Data API, fut interrogée avec la bibliothèque Pytube. Il est important de noter que YouTube impose une limite de 50 éléments pouvant être retournés par chaque requête. Pour contourner cette limite, la recherche a été découpée en plusieurs requêtes selon un critère de temps afin d'éviter les redondances en procédant par ordre décroissant de parution de la vidéo.

L'extraction des métadonnées des vidéos choisies a été réalisée au travers du module "YouTube" de la bibliothèque Pytube. Pour chaque lien vidéo nous avons extrait le titre, la description, le propriétaire de vidéo, le nombre de vues, la durée (en secondes), l'identifiant de chaîne, l'adresse web (url) de la chaine, les mots clés, la date de publication et l'identifiant unique de vidéo. Ces données ont été enregistrées dans un tableur en forma .csv pour les phases ultérieures de l'analyse.

En fin d'extraction, des analyses préliminaires ont permis de tirer des statistiques descriptives portant notamment sur une synthèse des métadonnées et un classement quantitatif par vidéo, par chaîne, etc. De plus, les vidéos ont été rangées et archivées en vue des prétraitements.

2.2 - Prétraitements de la base de données

Avant d'analyser les données et des métadonnées, certains prétraitements se sont montrés nécessaires. Le plus important a été l'extraction des transcriptions, ensuite traitées pour obtenir un corpus de contenu harmonisé.

La transcription des vidéos a été obtenue via la fonctionnalité « YouTube Transcript API » (Lawson 2015 ; Baudet and Point 2017 ; Zeni, Miorandi, and De Pellegrini 2013 ; Bachubhay et al. 2021 ; Alrashed et al. 2020 ; Paredes-Valverde et al. 2012). À partir de l'identifiant unique de chaque vidéo, cette fonctionnalité a permis d'extraire les transcriptions des vidéos ayant un sous-titrage.

L'extraction des transcriptions a mis en évidence une série de limites de la collecte : une partie des transcriptions étaient en langues différentes du français. De plus, la ponctuation était absente et tous les texte étaient en minuscule. Et finalement, plusieurs mots générés automatiquement,

ainsi que la présence de symboles et émoticônes introduisaient une source de bruit dans l'analyse. En conséquence, les textes ont dû être standardisés en fonction de ces différents points.

Les textes non français ont été traduits à l'aide de la bibliothèque python Translate (Derfoufi 2019 ; Liu 2020). Toutefois, la traduction automatique n'a pas pu gérer les caractères spéciaux. De plus, dans notre configuration d'utilisation, la bibliothèque limitait la traduction à 1000 signes (Kaleb 2022 ; Research 2022), et la plupart des transcriptions étaient des textes bruts n'ayant pas de délimiteurs. Il a donc été nécessaire d'affiner le prétraitement en effectuant une traduction ligne par ligne.

Une fois le texte transcrit, traduit et homogène, il est nécessaire de le normaliser. L'étape de normalisation permet une efficacité du texte dans la recherche d'informations. Pour cela, des algorithmes de traitement de langage naturel (NLP) sont appliqués. NLP est une branche de la science des données permettant aux machines de comprendre, manipuler, traduire ou générer le langage naturel, c'est à dire compréhensible par les humains tel qu'il est parlé et/ou écrit. Le NLP se positionne ainsi à l'intersection entre l'informatique, l'apprentissage automatique (machine learning) et la linguistique (Cambria and White 2014 ; Sag et al. 2002).

L'harmonisation et la standardisation des transcriptions a été réalisée à l'aide de deux bibliothèques python : SpaCy, dédiée au traitement de grands volumes de textes (Al Omran and Treude 2017 ; Srinivasa-Desikan 2018 ; JUGRAN et al. 2021 ; Altinok 2021 ; Vasiliev 2020) et NLTK, permettant la création d'interfaces avec plus de 50 ressources lexicales (e.g. WordNet) et des bibliothèques de classification du texte, ainsi que la tokenisation, l'analyse syntaxique et le raisonnement sémantique (Wang and Hu 2021 ; Hardeniya 2015 ; Hardeniya et al. 2016 ; Millstein 2020 ; Yogish, Manjunath, and Hegadi 2019 ; Farkiya et al. 2015). Les différentes manipulations des textes correspondent aux opérations classiques dans le traitement automatique du langage naturel. Ces opérations portent sur le nettoyage et la normalisation afin de réduire la taille du vocabulaire et, donc, de diminuer la dimension de l'espace d'entrée dans les modèles d'apprentissage automatique.

Le nettoyage du corpus des transcriptions comporte la suppression d'éléments inutiles et de mots vides (stopwords en anglais). Les éléments inutiles sont les onomatopées produites par la transcription

automatique (e.g. ehhh, ouhhh), ainsi que les émoticônes et d'autres symboles indésirables. Les mots vides sont des mots des expressions intercalaires régulières qui ont un pouvoir discriminant assez faible, tels que : et, ou, donc, la, les. En plus des mots vides du langage courant français, nous avons supprimé les mots vides spécifiques aux vidéos YouTube, tels que : bonjour, épisode, notification, abonnez, salut, etc. Ce nettoyage a permis de réduire la dimensionnalité du corpus et donc de focaliser le vocabulaire sur les mots discriminants pour le sujet d'étude. Le corpus nettoyé a été ensuite divisé en entités discrètes (tokens en anglais) par le découpage en mots individuels. À ce niveau nous avons supprimé la ponctuation (i.e. tokens qui sont des signes de ponctuation).

En utilisant la bibliothèque NLTK, nous avons ensuite passé tout le texte en minuscules, car la modélisation y est sensible, et supprimé espaces multiples et retours à la ligne car inutiles à la modélisation. Finalement, le corpus ainsi prétraité a été lemmatisé : en utilisant un dictionnaire et en tenant compte du contexte dans la phrase, chaque mot (i.e., token) a été reconduit à la racine, c'est-à-dire à sa forme neutre canonique. Le résultat a été converti en fichier « pickle » pour rendre le programme exécutable sur python.

2.3 - Modélisation des sujets

La modélisation des sujets (topic modeling en anglais) porte sur l'utilisation d'un modèle probabiliste afin d'identifier les thèmes qui peuvent décrire et résumer le contenu d'un texte opportunément prétraité. Il s'agit d'une méthode de classification non supervisée de document. Cette modélisation est considérée comme un type spécifique de clustering (regroupement) qui permet de détecter les sujets abordés dans un corpus de documents, assigner les sujets détectés à ces différents documents (Fedushko 2016 ; Li and Lei 2021 ; Nikolenko, Koltcov, and Koltsova 2017). La modélisation des sujets fournit des méthodes pour organiser, comprendre, rechercher et résumer automatiquement un ensemble de documents parlant de plusieurs thèmes (Brookes and McEnery 2019).

	LDA	Top2vec	BerTopic
Définition	Bibliothèque qui utilise des a priori Dirichlet pour les distributions document-sujet et mot-sujet, se prêtant à une meilleure généralisation.	C'est un algorithme de recherche sémantique. Il détecte automatiquement les sujets présents dans le texte et génère des vecteurs de sujet, de document et de mots intégrés	Bibliothèque qui utilise des transformateurs et TF-IDF (Term Frequency-Inverse) (Document Frequency) basé sur des classes pour créer des clusters denses
Avantages	La plus populaire et la plus efficace	Elle exploite l'incorporation sémantique conjointe de documents et des mots pour trouver des vecteurs thématiques. Pas besoin de lemmatisation, le nombre de sujets est trouvé automatiquement.	Elle interprète et visualise facilement les sujets générés. Flexible avec plusieurs langues
Inconvénients	Besoin du nombre de sujets à connaître. Ignore l'ordre et la sémantique des mots	Ne contient pas de manière de visualisation graphique	Beaucoup de documents classés aberrants

Tableau 2. Comparison des trois bibliothèques les plus fréquentes pour une modélisation de sujet thématique (Fedushko 2016 ; Bhat et al. 2020 ; Alsumait et al. 2010 ; Xie et al. 2020 ; Egger and Yu 2022)

Des trois bibliothèques utilisées plus fréquemment pour la modélisation des sujets (Tableau 2), nous avons exclu LDA car il nécessite de fournir à l'avance le nombre de sujets. Il aura aussi fallu réaliser plusieurs itérations pour en régler les hyperparamètres permettant de créer des thèmes significatifs. Par conséquence, seule les bibliothèques top2vec et BerTopic ont été comparées pour modéliser le corpus des transcriptions.

2.3.1 - Top2vec

L'algorithme part du principe que l'existence de nombreux documents

sémantiquement similaires est révélatrice d'un sujet sous-jacent. La première étape consiste à créer un ensemble de vecteurs de documents et de mots. Une fois les documents et les mots intégrés dans un espace vectoriel, l'objectif de l'algorithme est de trouver des groupes de documents denses, puis d'identifier les mots qui ont rassemblé ces documents. Chaque zone dense est un sujet et les mots qui ont attiré les documents dans la zone dense sont les mots du sujet. Une fois le modèle Top2Vec entrainé, nous avons obtenu le nombre de sujets de notre corpus et exploité ses différentes fonctionnalités. Le modèle, une fois entraîné, détecte automatiquement les sujets présents dans l'ensemble de corpus (2 dans notre cas) et génère un vecteur pour chaque sujet, les documents qui le définissent ainsi que les mots intégrés qui le caractérisent avec leurs scores respectifs. Les 50 premiers mots et les scores de similarité sont renvoyés, par ordre de similarité sémantique avec le sujet.

2.3.2 - BerTopic

Intégration des documents (embedded documents en anglais). Nous commençons par créer des incorporations de documents à partir d'un ensemble de documents en utilisant des transformateurs de phrases. Ces modèles sont pré-entraînés pour de nombreuses langues et sont parfaits pour créer des incorporations de documents ou de phrases.

Regroupement des documents. Nous réduisons la dimensionnalité des incorporations que nous avons générés via la bibliothèque UMAP. Ensuite, nous utilisons la bibliothèque de clustering basée sur la densité (HDBSCAN) pour créer nos clusters et identifier les aberrations.

Création de la représentation du sujet. Nous utilisons la bibliothèque TF-IDF pour savoir le contenu de chaque cluster généré. La bibliothèque TF-IDF permet de trouver des mots intéressants par groupe de documents plutôt que par chaque document. En appliquant le TF-IDF sur l'ensemble de documents, nous comparons l'importance des mots entre les documents. Ainsi, tous les documents d'une même catégorie sont traités comme un seul document sur lequel TF-IDF est ensuite appliqué. Les résultats sont des scores des poids pour les mots au sein d'un cluster. Plus les mots sont significatifs dans un cluster, plus ils sont représentatifs de ce sujet. Chaque cluster est converti en un seul document au lieu d'un ensemble de documents.

Cohérence de la pertinence marginale maximale. Avec les représentations TF-IDF, nous avons un ensemble de mots qui décrivent une collection de documents. Techniquement, cela ne signifie pas que cette collection de mots décrit un sujet cohérent. En pratique, nous verrons que de nombreux mots décrivent un sujet similaire, mais que certains mots seront, d'une certaine manière, trop adaptés aux documents. Pour améliorer la cohérence des mots, la pertinence marginale maximale a été utilisée pour trouver les mots les plus cohérents sans avoir trop de chevauchement entre les mots eux-mêmes. Cela permet d'éliminer les mots qui ne contribuent pas à au sujet.

2.3.3 - Regroupement par mot clé

À partir de ces modèles il a été possible d'identifier une liste des agroéquipements les plus traités par les vidéos YouTube. Cependant, d'autres machines agricoles ne sont pas considérées comme mots significatifs pour le modèle. Par conséquent, il n'a pas été possible de les regrouper. Pour remédier à ce problème, nous avons classé les documents des mots non traités selon le principe suivant : un document qui contient plus d'occurrences d'un mot clé est un document significatif et qui parle sûrement de cette machine agricole. C'est à travers la méthode "search" de la bibliothèque "regexp" que les documents les plus pertinents du corpus des transcriptions ont été extraits, et la pertinence ici est définie par la fréquence d'apparition du terme en question.

3 - Résultats

Les résultats présentés dans cette section sont donc le résultat d'un outil exploratoire. L'analyse de ces résultats permettront donc aussi de valider la méthode.

3.1 - Collecte des données et description du corpus

Au total, 311 liens et par conséquent autant de vidéos couvrant la période 2005-2022 sont obtenus (Tableau 3). Les pratiques les plus traitées par les vidéos sont la récolte (N = 139) et le « travail du sol » (N =93). On note également que la dernière année (2021 jusqu'à août 2022) présente le nombre le plus important des vidéos publiées sur la période couverte.

	2005-2011	2011-2013	2013-2015	2015-2017	2017	2018	2019	2020	2021-2022	Total
Travail du sol	1	2	3	2	2	6	1	16	60	93
Semis	0	1	0	0	1	9	6	7	24	48
Désherbage	0	0	1	4	0	0	0	1	2	8
Récolte	0	1	0	2	4	5	4	24	96	139
Ensilage	0	0	0	0	0	5	4	1	13	23
Total	1	4	4	8	10	25	15	49	195	311

Tableau 3. Résultat du scraping des liens. Nombre brute de vidéos par année

La suppression des redondances mène à 234 liens YouTube uniques. Dix liens YouTube russes ont été bloqués à cause de la guerre en Ukraine. Les liens YouTube disponibles sont donc 224. Ces liens ont été archivés dans un fichier de type bloc-notes (.txt). Les métadonnées associées concernant plusieurs descripteurs ont été archivés dans un fichier .csv.

L'analyse des métadonnées a fait émerger un inconvénient du data API de YouTube qui restitue des liens hors domaine. Un filtrage des liens a donc été effectué ou seules les vidéos ayant un mot du dictionnaire [cf tableau 1 pour les mots clés] dans le titre ola u description ont été gardés. Après filtrage, une liste de 176 liens a été obtenue. Il est constaté que la méthode extrait plus de contenu que nécessaire nécessitant plusieurs tris.

Ces vidéos sont repartis entre 68 chaines YouTube. Pour chaque chaîne nous avons documenté : nom de la chaîne, nombre d'abonnés, nombre de vues, identifiant unique, nombre total de vidéos, identifiant de la playlist (Figure 3).

	Channel_name	Subscribers	Views	channel_id	Total_videos	playlist_id
0	PowerBoost	58600	23765057	UCXur-8uR91-dEiafZevXwGQ	1519	UUXur-8uR91-dEiafZevXwGQ
1	Jilles Boer	12600	2865421	UCZdAJQWIhuLkAx0VZ2mMaFg	148	UUZdAJQWIhuLkAx0VZ2mMaFg
2	Biomanie	3650	519208	UCW39-vjpoz7-RuJDaiZSE8A	32	UUW39-vjpoz7-RuJDaiZSE8A
3	Stauffer Garage	1290000	194699046	UC6f1N1Cg__5Ex0aWBUZeYGg	219	UU6f1N1Cg__5Ex0aWBUZeYGg
4	FieldView	3400	14451945	UCegW3qxC77Yrg-nKH-VUMFA	295	UUegW3qxC77Yrg-nKH-VUMFA
...	...	...	...	...	...	...
15	Ricardos Gaming	12500	4615121	UCYTr93ZqkpdyDh_GTe39bdA	1569	UUYTr93ZqkpdyDh_GTe39bdA

Figure 3. Tableau des informations des chaînes

4 - Prétraitement des données

À la suite du processus de traduction des transcriptions, nous remarquons que la majorité des vidéos disposait d'une transcription en français et que la totalité des vidéos issues du filtrage a pu être traitée. Au total nous pouvons construire un corpus de 176 transcriptions des vidéos (Tableau 4).

Vidéo	Nombre de liens
Avec transcription en français	133
Avec transcription (< 10000 caractères) traduite en français	18
Avec transcription >10000 caractères	23

Tableau 4. Résultats de la traduction des transcriptions (N = 176)

Les résultats du prétraitement de transcription ont engendré des textes nettoyés, prétraités, lemmatisés, prêts à l'exploration et à la modélisation. Ces résultats ont été archivés dans des fichiers du type bloc-notes.
Modélisation des données

Nous allons passer en revue les principaux composants de Top2Vec et BerTopic et les étapes nécessaires pour créer des modèles thématiques solides.

4.1 - Première application : Top2vec

L'entraînement du modèle se fait avec la fonction suivante :
model = Top2Vec (document= corpus, speed=« learn », workers=8)

Les résultats montrent que le lexique utilisé par les agri-youtubeurs est composé de verbes d'action qui reflètent le terme bricolage.

Exemple d'utilisation de Top2Vec : Si on prend à titre d'exemple le mot 'semoir', on cherche les sujets les plus similaires à ce mot via un score pouvant aller de 0 (le plus similaire) à 1 (le moins similaire). Il possible alors de générer un nuage de mots (Figure 4. Le nuage de mot les plus similaire au semoir dans topic 1) en relation avec le mot choisi qui est 'semoir' dans ce cas.

Figure 4. Le nuage de mot les plus similaire au semoir dans topic 1

Les documents ayant un score supérieur à 0.5 (seuil de choix) ont été classés dans la thématique 'agroéquipement'. Les machines agricoles traitées et reconnues par le modèle sont : charrue, herse, moissonneuse-batteuse, semoir, presse et tracteur. Il est possible de constater que la plupart les vidéos illustrent effectivement ces agroéquipements.

4.2 - Deuxième application : BerTopic

Ici, nous explorons les fonctionnalités les plus fréquentes du modèle BerTopic pour créer des catégories à partir du corpus de données. Pour pouvoir entraîner le modèle BerTopic, une activation du GPU, et une installation de la bibliothèque BerTopic sont obligatoires.

Extraction des sujets. Après avoir ajusté le modèle, il est possible d'examiner les résultats. Nous regardons d'abord les sujets les plus fréquents car ils représentent mieux le corpus de mots. Le sujet le plus fréquent (sujet 0) regroupe 62 documents, dont les mots les plus représentatifs indiquent un thème agricole en lien avec la ferme et le semis (Figure 5).

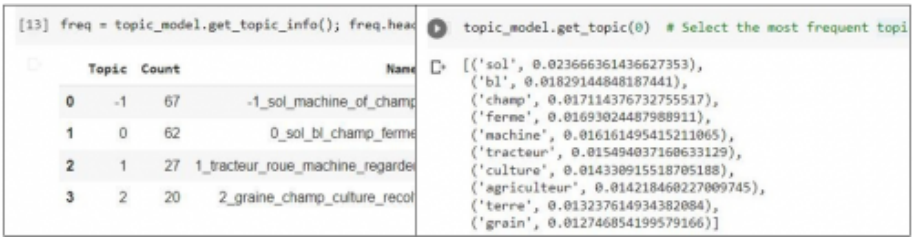

Figure 5. Exemple de résultat modèle issu de BerTopic : à gauche, liste des sujets ; à droite, mots représentatifs du sujet le plus fréquent (topic 0) au moins fréquent (topic 3). Le topic -1 fait référence à toutes les valeurs aberrantes qui doivent être ignorées.

Plusieurs options de visualisation sont disponibles dans BerTopic, notamment la visualisation des termes des sujets, des probabilités des matrices de similarité. Par exemple, il est possible de visualiser les termes pertinents pour chaque sujet en créant des diagrammes à barres à partir des scores c-TF-IDF pour chaque représentation de sujet. Ces scores permettent d'indiquer la signification de chaque sujet et de comparer les représentations des sujets entre elles (Figure 6).

Figure 6. Scores enregistrés pour les mots les plus significatifs des thèmes

La visualisation des probabilités des sujets sert à définir la distribution des sujets dans le corpus de données : le deuxième sujet est le plus probabiliste avec une distribution de 0.4. Une matrice de similarité permet de calculer la corrélation entre les différents sujets. (Figure 7).

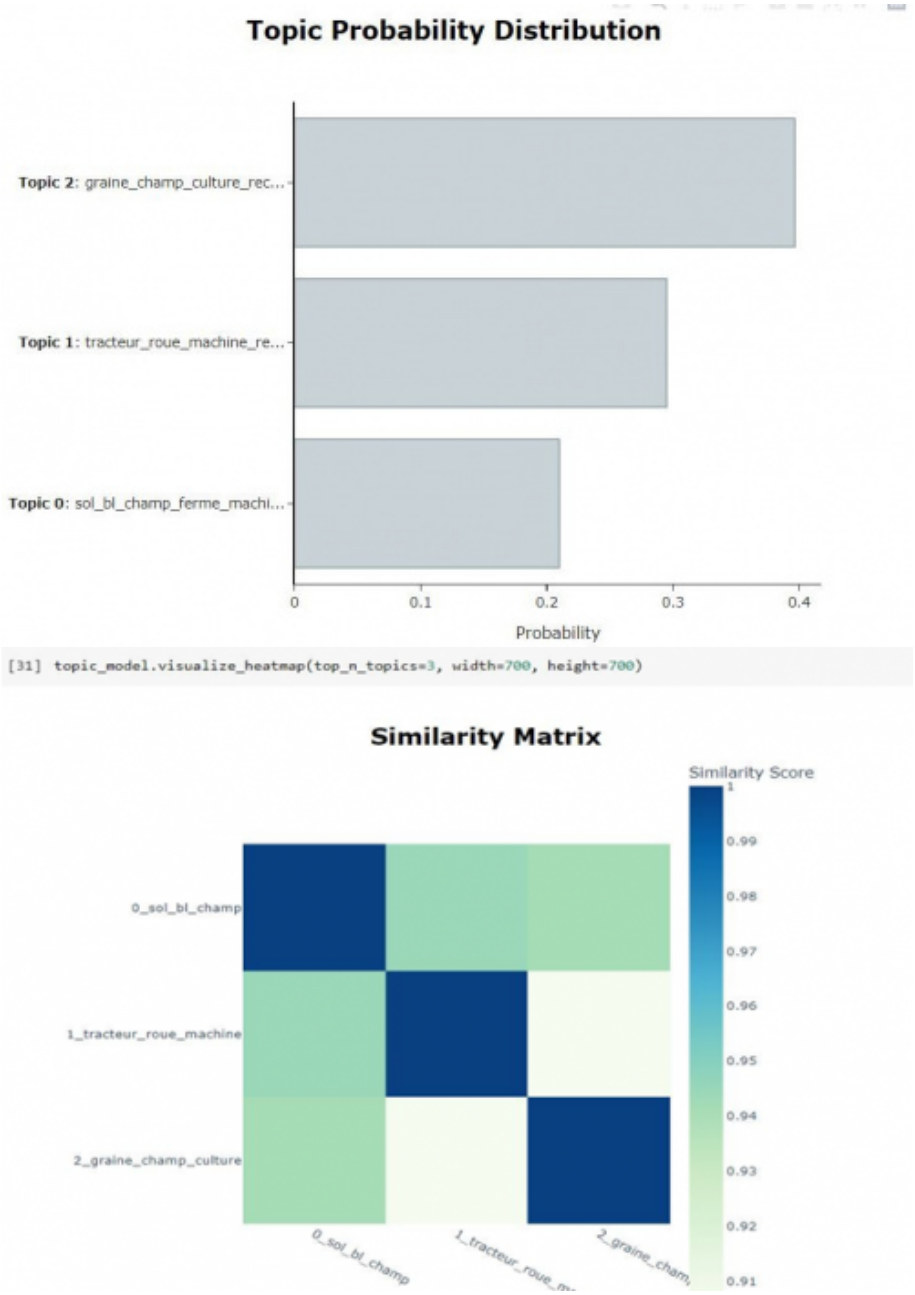

Figure 7. Analyse du classement des sujets par BerTopic. En haut, distributions des documents. En bas, matrice de similarité entre les sujets ; les sujets qui ont de fortes ressemblances sont : les sujets 1 et 0 avec un score de 0.95

Après avoir créé le modèle thématique, nous avons remarqué qu'il y a deux sujets qui sont presque semblables (sujet 0 et sujet 2). Il est possible de les considérer comme un seul. BERTopic permet de réduire le nombre de sujets. Le premier sujet étant aberrant, nous sommes passés de 3 sujets à 2 sujets. Le premier sujet (sujet 0, Figure 5) parle de tout ce

qui est en rapport avec le sol et les cultures. Le deuxième sujet (sujet 1, Figure 5) aborde directement les thèmes de machinisme agricole et d'agroéquipements.

4.3 - Structuration des résultats

La démarche porte principalement sur l'identification des documents les plus pertinents en relation avec les machines agricoles en vue d'extraire les bricolages de ces derniers. Les bricolages sont exposés dans les vidéos YouTube par les agri-youtubeurs ou les youtubeurs d'une manière très explicite. Voici quelques cas de figures par lesquelles ils expriment les bricolages :
– L'agriculteur parle de son agroéquipement, explique son fonctionnement en donnant indirectement des indicateurs d'une bonne exploitation de l'ensemble de composants de cet équipement.
– Les constructeurs des machines détaillent les nouveautés et l'utilité en rapport avec un besoin spécifique.
– Un agriculteur partage son expérience en exerçant sur son champ une pratique (récolte ou semis, ou désherbage...) sur une culture.
– Un agriculteur rencontre un problème durant la récolte et essaye de les résoudre par le moins de dégâts en exposant les causes, les conséquences et les différentes alternatives pour ne pas les réaffronter.
– Un agriculteur déploie deux ou plusieurs agroéquipements pour atteindre un objectif ou satisfaire un besoin.
– Un agriculteur et son fils parlent de l'exploitation familiale, ainsi que de leur semoir trip-till (cultivation en bandes) unique qu'ils ont fabriqué eux-mêmes.

À ce stade, des groupes qui classent chaque agroéquipement avec le top 5 des documents les plus pertinents ont été construits. Un obstacle important est le manque d'une base de données standard permettant d'entraîner les modèles et extraire ensuite les bricolages explicites de chaque machine agricole de manière automatique. Cette étape est donc réalisée manuellement. Cela donne un ensemble de données structuré de trois colonnes : bricolage, agroéquipement et besoins.

5 - Discussions

Concernant la démarche analytique, plusieurs constats sont possibles tant sur les innovations et limites de la démarche proposée que sur les résultats trouvés.

5.1 - Innovations et limites de la démarche

La méthodologie appliquée dans notre étude a été modulée et a évoluée au fur et à mesure de l'avancée du processus de recherche : nous avons géré, une par une, les contraintes émergentes relatives à l'outil d'extraction de données (YouTube) et à l'analyse de la base de données construites. Cette approche du bricolage, ici par l'usage de technologies, a permis de respecter la complexité des processus de construction des résultats, notre méthode est donc elle-même, du bricolage (Denzin and Lincoln 2008). Denzin et Lincoln (2008) suggèrent que la combinaison de multiples pratiques méthodologiques, de matériaux empiriques, dans une même étude, ce qui relève du bricolage, doit être comprise comme une stratégie qui ajoute de l'ampleur, de l'efficacité, de la complexité, de la richesse et de la profondeur à l'étude, ce en quoi ils confirment la créativité liée au bricolage.

Nous constatons deux innovations (I) et trois limites (L) principales à cette méthodologie :
– (I) La collecte des transcriptions est faite à partir d'une source d'information inexplorée dans les processus de construction des corpus textuels dans les tâches de fouille de textes, à savoir YouTube. Cette source d'information s'est avérée riche en contenu (les vidéos YouTube) pour décrire les expériences des agriculteurs autour des bricolages des agroéquipements, bien que davantage pertinente pour l'apprentissage visuel.
– (I) Quand il est question d'utilisation du démonstratif dans les différents discours (ex. « Je fais ça » « Je change ceci » …), l'agriculteur ne nomme pas le bricolage, il le montre. Des algorithmes de segmentation d'image poussés comme 'Segment anything' et de machines d'apprentissage multimodaux pourront analyser ce type de données
– (L) Le tableau des pratiques agricoles utilisé pour faire l'extraction de vidéos peut ne pas être exhaustif. En effet, les pratiques agricoles et les agroéquipements pris en compte sont les pratiques et les agroéquipements les plus communs et les plus basiques. Il est possible que les agriculteurs les plus bricoleurs ne s'exposent pas sur YouTube ou sur des canaux spécifiques. Pour améliorer la collecte de données, il peut être très utile d'analyser de plus près le point d'entrée dit "playlist".Enfin, pour ce qui concerne les méthodes pour la modélisation thématique des sujets, BerTopic et Top2vec sont deux modèles stochastiques : les sujets peuvent ainsi différer d'une exécution à l'autre.
– (L) Les limites dérivent de la multiplicité de termes technique et

éléments de langage employés dans les vidéos, que la traduction automatique n'arrive pas à rendre correctement quand s'agit de langage informel ou de langage familier. De plus, les hésitations à l'oral typique des vidéos improvisées dégradent la compréhensibilité des mots, souvent coupés, manquants ou répétés. Enfin, l'extraction de transcription de vidéos implique l'incomplétude des descriptions, remplacées par des termes démonstratifs et des indications de lieux appauvrissant le traitement automatique des contenus. Sur le côté analytique, le traitement automatique fait perdre ou ne rend pas compte de la division en phrases. Les mots peuvent donc perdre le contexte à cause d'une réorganisation aléatoire des phrases.
– (L) Concernant les outils analytiques choisis, il faut considérer que la fonctionnalité "search" de l'API YouTube renvoie un ensemble des résultats de recherche qui correspondent aux critères définis pour identifier les ressources vidéo, les chaînes et bien d'autres informations. Malgré son efficacité, cette méthode présente des restrictions : la contrainte de quotas limite les résultats extraits. Par défaut, les projets qui activent l'API de données YouTube dispose seulement d'un quota de 10 000 unités par jour, dont chaque méthode API a une unité de quota. La fonctionnalité search.list est donc très chronophage.

5.2 - Portée et limites de l'échantillon de vidéos

Les métadonnées descriptives ne permettent pas de faire des analyses en Big Data car l'échantillon est petit. Elles permettent néanmoins d'avoir les tendances sur les vidéos publiées. Nous constatons à ce propos que les publieurs de vidéos ne sont pas uniquement des agri-youtubeurs. Nous trouvons des constructeurs d'équipements, des instituts de recherche et des journalistes qui publient des vidéos sur le bricolage en agroéquipements, alors que cela ne donne pas lieu à des articles construits, car ce sont des vidéos commerciales.

L'augmentation du nombre de publications sur ce sujet par YouTube durant les années 2020, 2021 et 2022 est également constatée. Cela peut être associé à différentes raisons : une tendance générale du nombre de vidéos publiées chaque année, une plus grande utilisation des réseaux sociaux par les agriculteurs et un intérêt croissant ces trois dernières années pour le bricolage des agroéquipements. Si nous supposons que les agriculteurs de cet échantillon manifestent un plus grand intérêt à bricoler leurs agroéquipements, nous pourrons en déduire que leurs agroéquipements ne sont pas bien adaptés à leurs besoins. Les

tendances à l'augmentation des bricolages de l'agroéquipement peuvent également être un indicateur de la tendance à la non-adaptation du parc matériel avec les besoins des agriculteurs lesquels seraient en train de changer avec la nécessité d'agir pour la transition vers l'agroécologie.

Différentes questions se posent. Quelles sont donc les besoins qui poussent les agriculteurs à bricoler leurs agroéquipements ? Les raisons peuvent être économiques, agronomiques, environnementales, en lien avec le dérèglement climatique, ou tous ces éléments réunis. Savoir quels agroéquipements est bricolés et pourquoi ils le sont, pourrait répondre à ces questions.

5.3 - Intérêts et limites des sujets modélisés

Après analyse des métadonnées descriptives, la modélisation des données textuelles via le modèle BerTopic a permis d'extraire les deux sujets les plus significatifs dans le corpus de mots des transcriptions. « Sol, champs » est le sujet le plus significatif dans les vidéos publiées. Le sol et le champ sont donc fréquemment mentionnés quand il est question de bricolage en agroéquipement.

Le sujet « sol, champs », fréquemment abordé par les agriculteurs, pourrait expliquer les problématiques liées à la santé du sol et du champ et sa préservation (par la préservation de sa biodiversité, de la couche arable du sol, du taux de matière organique) contre les potentielles agressions (érosion, labour profond, appauvrissement en matière organique et en biodiversité).

Il possible de conclure que pour cet échantillon de vidéos YouTube, le sujet « sol, champ » suscite des bricolages d'agroéquipements, cela pourrait indiquer que les agroéquipements concernés ne sont pas suffisamment adaptés.

Pour ce qui est des agroéquipements bricolés, nous constatons que les agroéquipements qui ont suscité le plus de bricolages sont le semoir et la moissonneuse batteuse, touchant les pratiques agricoles les plus cruciales, de manière générale, et plus spécifiquement dans l'adaptation vers la transition agroécologique.

Pour aller plus loin, il y a trois types de voies possibles. La première serait de voir et écouter en totalité les 176 vidéos. On sait qu'elles sont triées et

montreront des données intéressantes de bricolages montrés et non nommés. La deuxième, nous l'avons noté en 4.1, serait d'utiliser des algorithmes de segmentation d'image poussés. La troisième voie serait de réaliser des entretiens semi-directifs d'agriculteurs triés sur le fait qu'ils bricolent pour savoir de quoi il s'agit exactement.

5.4 - Perspectives générales sur le bricolage

Parallèlement à Lévi-Strauss et à l'intérêt anthropologique qu'il porte au concept du bricolage, le généticien Jacob remet en question l'analogie faite et validée par la littérature antérieure entre le processus d'évolution des espèces par sélection naturelle (Darwin and Darwin 1964) et les méthodes d'ingénierie telles que la planification, le design du modèle avant sa fabrication et l'importance de l'exactitude dans l'exécution (Bendall 1983). Jacob (1977) associe le processus de l'évolution des espèces au bricolage. Expliqué par Racine (2014), Jacob « soutenait que les effets cumulatifs de l'histoire sur l'évolution de la vie, mis en évidence par les données moléculaires, fournissaient une autre explication des modèles décrivant l'histoire de la vie sur terre. » Repenser le processus de l'évolution par la sélection naturelle selon le concept de bricolage a permis d'approfondir les recherches en biologie de l'évolution et du développement sur la régulation génétique, les évènements de duplication de gènes et le programme génétique du développement embryonnaire (2014). En introduisant le concept de bricolage au public scientifique, Jacob s'appuie sur des arguments de biologie moléculaire et de l'évolution (Laubichler 2006). Le bricolage de Jacob serait donc défini comme étant « des abstractions conceptuelles qui permettent l'analyse théorique d'un large éventail de phénomènes qui sont unis par un processus sous-jacent commun (le bricolage) ou le réarrangement et la recombinaison opportunistes d'éléments existants. ». Par ailleurs, Jacob soutient que les travaux de recherche en biologie vont rencontrer de nombreuses imperfections dans l'histoire de l'évolution ainsi que « des similarités sous-jacentes dans les structures moléculaires des génomes malgré l'influence de la sélection génétique ». Nous constatons ici une convergence entre deux pensées issues de champs disciplinaires différents (anthropologie et biologie) : le concept de bricolage est nécessaire pour comprendre différents processus évolutifs.

Créer un nouveau concept n'est pas facile ; il pourrait paraître souhaitable de le définir sans tautologie ou sans utiliser des concepts indéfinissables ou inconcevables. Selon Francis Kaplan, « La seule

preuve qu'on puisse donner que la tâche est possible, c'est de la réaliser. Or, certains concepts sont indéfinissables comme la conscience, la vie, l'espace et le temps. […] On ne peut analyser qu'avec des concepts ; on ne peut analyser des concepts qu'avec des concepts qui en permettent l'analyse et il n'existe pas de concepts capables d'analyser des concepts fondamentaux – c'est même ce qui définit ceux-ci. […] Ce qui est vrai de l'évolution est aussi vrai de l'esprit. Lui aussi procède par bricolage ; il se sert d'un concept destiné à un usage pour un usage différent. […] La compréhension de la vie relève donc du bricolage de l'esprit. Certes comme tout bricolage, ce bricolage est peu satisfaisant sur le plan intellectuel ; mais comme tout bricolage aussi, il est efficace, comme le montrent les succès de la biologie » (Kaplan 1994, p 238-259).

Francis Kaplan ira plus loin dix ans plus tard, en étudiant les concepts bricolés (Kaplan 2004, p 81-181). En quelque sorte le bricolage de l'esprit qui nous permet de concevoir le bricolage, de manière générale, est un concept fondamental. « En fait, on bricole des concepts comme le bricoleur bricole des outils dont il ne comprend pas le fonctionnement, mais qui fonctionnent. […] Il est, par conséquent, tout à fait légitime de dire que la force d'attraction est un concept bricolé, parce que non concevable, mais efficace ». […] Les concepts bricolés n'ont pour rôle que de nous faire comprendre. Sans doute leur explication est incohérente, donc insuffisante ; elle n'en est pas moins réelle […] et il n'est pas contraire à l'essence d'un concept bricolé efficace de pouvoir être remplacé par un autre concept bricolé plus efficace (Kaplan 2004 p 98). Ainsi, en physique, le concept de champ remplacera celui d'attraction.

De nombreux concepts fondamentaux, y compris en science, sont de l'ordre du bricolage de l'esprit. Il est relativement facile de lire de la philosophie, mais il est plus difficile de philosopher ; relativement facile d'apprendre les acquis des sciences et plus difficile d'élargir le champ des connaissances scientifiques. De même il est relativement facile d'observer et d'analyser comment une personne escalade une paroi, mais il est plus difficile d'escalader soi-même ; il est, finalement, relativement facile d'analyser le fonctionnement du bricolage technique, mais il est plus difficile de réaliser un bricolage technique. Il en ressort que le concept de bricolage est un concept fondamental. Qu'il s'agisse de technique, d'ingénierie, de science ou de philosophie, le bricolage est une sorte de nécessité pour leur permettre de progresser. Dans ces trois domaines, le bricolage est au cœur de la créativité et de l'inventivité des

humains. Nous venons de constater que face aux changements nécessaires de l'agriculture, les agriculteurs qui innovent, bricolent.

6 - Conclusions

Cette recherche méthodologique sur le bricolage en tant qu'approche qualitative (Hammersley 2004) montre qu'il est possible de mesurer une potentielle inadaptation des agroéquipements par rapport au parc matériel actuel dans le cadre d'un changement d'approche fondamental : dans le cas des agroéquipements, le bricolage des agriculteurs est un exemple concret et très actuel de créativité et d'inventivité humaine. En étant le plus efficace possible, tant dans la façon de faire que dans l'objectif, l'agriculteur use de son ingéniosité, de sa technique et des outils disponibles pour adapter son agroéquipement à ses besoins (agronomiques, financiers etc.). Bricoler l'agroéquipement exprime un besoin de l'adapter, d'améliorer ou de le faire gagner en efficacité, néanmoins d'autres besoins peuvent ressortir suite à ce bricolage. Il sera intéressant de vérifier comment est-ce que les bricolages des agriculteurs permettront cette adaptation et ce gain en efficacité. En perspective, en référence aux trois voies proposées en 4.3, des entretiens ciblés semi-directifs devraient y être adaptés. La méthodologie de l'étude actuelle basée sur la fouille de données YouTube se base sur l'exploration et la description des tendances de publication et permet de confirmer l'intérêt de discuter plus en détails avec des agriculteurs en agriculture de conservation des sols.

Bibliographie

Agarwal, Neha, Rajat Gupta, Sandeep Kumar Singh, and Vikas Saxena
2017 Metadata Based Multi-Labelling of YouTube Videos. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence Pp. 586–590.

Aggarwal, Nisha, Swati Agrawal, and Ashish Sureka
2014 Mining YouTube Metadata for Detecting Privacy Invading Harassment and Misdemeanor Videos. In 2014 Twelfth Annual International Conference on Privacy, Security and Trust Pp. 84–93.

Al Omran, Fouad Nasser A, and Christoph Treude
2017 Choosing an NLP Library for Analyzing Software Documentation : A Systematic Literature Review and a Series of Experiments. In 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR) Pp. 187–197.

Alrashed, Tarfah, Jumana Almahmoud, Amy X. Zhang, and David R. Karger
2020 ScrAPIr : Making Web Data APIs Accessible to End Users. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems Pp. 1–12. CHI '20. New York, NY, USA : Association for Computing Machinery. https://doi.org/10.1145/3313831.3376691, accessed November 9, 2022.

Alsumait, Loulwah, Pu Wang, Carlotta Domeniconi, and Daniel Barbara
2010 Embedding Semantics in LDA Topic Models. In Pp. 183–204.

Altinok, Duygu
2021 Mastering SpaCy : An End-to-End Practical Guide to Implementing NLP Applications Using the Python Ecosystem. Packt Publishing Ltd.

Archer, Geoffrey R, Ted Baker, and Rene Mauer
2009 Towards an Alternative Theory of Entrepreneurial Success : Integrating Bricolage, Effectuation and Improvisation. Frontiers of Entrepreneurship Research 29(6) : 16.

Bachubhay, Akhil, Danny Chhour, Heji Deng, and Trung Tran
2021 YouTube Video Analysis. Virginia Tech. https://vtechworks.lib.vt.edu/handle/10919/103255, accessed November 9, 2022.

Baudet, Cédric, and Sébastien Point, eds.
2017 Des données aux métadonnées de la société numérique : vers un codage 2.0. In Faire face à la complexité dans un monde numérisé P. 16. Paris, France : Association Information et Management (AIM). https://univ-lyon3.hal.science/hal-01519562.

Bendall, D. S.
1983 Evolution From Molecules to Men. CUP Archive.

Bentley, Jeffery W., Paul Van Mele, Nafissath Fousseni Barres, Florent Okry, and Jonas Wanvoeke
2019 Smallholders Download and Share Videos from the Internet to Learn about Sustainable Agriculture. International Journal of Agricultural Sustainability. World, Taylor & Francis. https://www-tandfonline-com.inc.bib.cnrs.fr/doi/abs/10.1080/14735903.2019.1567246, accessed September 1, 2022.

Bhat, Muzafar Rasool, Majid A Kundroo, Tanveer A Tarray, and Basant Agarwal
2020 Deep LDA : A New Way to Topic Model. Journal of Information and Optimization Sciences 41(3). Taylor & Francis : 823–834.

Brookes, Gavin, and Tony McEnery
2019 The Utility of Topic Modelling for Discourse Studies : A Critical Evaluation. Discourse Studies 21(1). SAGE Publications : 3–21.

Cambria, Erik, and Bebo White
2014 Jumping NLP Curves : A Review of Natural Language Processing Research [Review Article]. IEEE Computational Intelligence Magazine 9(2) : 48–57.

Casey, Sarah, Gail Crimmins, Laura Rodriguez Castro, and Penny Holliday
2022 "We Would Be Dead in the Water without Our Social Media !" : Women Using Entrepreneurial Bricolage to Mitigate Drought Impacts in Rural Australia. Community

Development 53(2). Routledge : 196–213.

Darwin, Professor Charles, and Charles Darwin
1964 On the Origin of Species : A Facsimile of the First Edition. Harvard University Press.

Denzin, Norman K., and Yvonna S. Lincoln
2008 Introduction : The Discipline and Practice of Qualitative Research. In Strategies of Qualitative Inquiry, 3rd Ed Pp. 1–43. Thousand Oaks, CA, US : Sage Publications, Inc.

,
eds.
2017 The SAGE Handbook of Qualitative Research. 5th edition. Thousand Oaks : SAGE Publications, Inc. https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-qualitative-research/book242504, accessed February 9, 2023.

Derfoufi, Younes
2019 Programmation En Langage Python. https://hal.archives-ouvertes.fr/hal-02126596, accessed November 9, 2022.

Egger, Roman, and Joanne Yu
2022 A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. Frontiers in Sociology 7 : 886498.

Farkiya, Alabhya, Prashant Saini, Shubham Sinha, and Sharmishta Desai
2015 Natural Language Processing Using NLTK and WordNet 6 : 6.

Fedushko, Solomia
2016 Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology (CCSEIT 2016), Vienna, Austria, May 21 22, 2016.

Glez-Peña, Daniel, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato, and Florentino Fdez-Riverola
2014 Web Scraping Technologies in an API World. Briefings in Bioinformatics 15(5) : 788–797.

Hammersley, Martyn
2004 Teaching Qualitative Methodology : Craft, Profession or Bricolage. In . Clive Seale, Giampietro Gobo, Jay F. Gubrium, and David Silverman, eds. Pp. 549–560. Thousand Oaks : Sage. http://www.sagepub.co.uk/booksProdDesc.nav?prodId=Book224935, accessed February 1, 2023.

Hardeniya, Nitin
2015 NLTK Essentials : Build Cool NLP and Machine Learning Applications Using NLTK and Other Python Libraries. Packt Open Source. Birmingham : Packt Publ.

Hardeniya, Nitin, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur
2016 Natural Language Processing : Python and NLTK. Packt Publishing Ltd.

Hatton, Elizabeth
1989 Lévi-Strauss's Bricolage and Theorizing Teachers' Work. Anthropology & Education

Quarterly 20(2) : 74–96.

JUGRAN, SWARANJALI, ASHISH KUMAR, BHUPENDRA SINGH TYAGI, and VIVEK ANAND
2021 Extractive Automatic Text Summarization Using SpaCy in Python & NLP. In 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) Pp. 582–585.

Jun Lee, Sang, and Keng Siau
2001 A Review of Data Mining Techniques. Industrial Management & Data Systems 101(1). MCB UP Ltd : 41–46.

Kaleb, Jordan
2022 How to Do Language Translation in Python. DEV Community. https://web.archive.org/web/20220202195427/https://dev.to/kalebu/how-to-do-language-translation-in-python-1ic6, accessed November 9, 2022.

Karthikeyan, T, Sekaran Karthik, D Ranjith, V Vinoth Kumar, and J. M Balajee
2019 Personalized Content Extraction and Text Classification Using Effective Web Scraping Techniques. International Journal of Web Portals (IJWP) 11(2). IGI Global : 41–52.

Khder, Moaiad
2021 Web Scraping or Web Crawling : State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and Its Applications 13(3) : 145–168.

Kincheloe, Joe L.
2011 Describing the Bricolage. In Key Works in Critical Pedagogy. Kecia Hayes, Shirley R. Steinberg, and Kenneth Tobin, eds. Pp. 177–189. Bold Visions in Educational Research. Rotterdam : SensePublishers. https://doi.org/10.1007/978-94-6091-397-6_15, accessed February 1, 2023.

La langue française
2022 Définition de bricoler. Dictionnaire français. https://www.lalanguefrancaise.com/dictionnaire/definition/bricoler, accessed February 9, 2023.

Laubichler, Manfred D.
2006 Tinkering : A Conceptual and Historical Evaluation. In Tinkering : The Microevolution of Development Pp. 20–34. John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470319390.ch2, accessed November 15, 2022.

Lawson, Richard
2015 Web Scraping with Python. Packt Publishing Ltd.

Lévi-Strauss, Claude
1962 La pensée sauvage. Agora, 2. Paris : Presses Pocket.

Li, Xin, and Lei Lei
2021 A Bibliometric Analysis of Topic Modelling Studies (2000–2017). Journal of Information Science 47(2). SAGE Publications Ltd : 161–175.

Liu, Hui
2020 A Study on the Reader Reception of English Translations of The Analects-Data Analysis with Python. In 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE) Pp. 2381–2387.

Marathe, Sunil Marathe, and Kavita P. Shirsat
2015 Approaches for Mining YouTube Videos Metadata in Cyber Bullying Detection. International Journal of Engineering Research & Technology 4(05) : 680–684.

Mehta, Kunal, Maya Salvi, Rumil Dand, Vineet Makharia, and Prachi Natu
2020 A Comparative Study of Various Approaches to Adaptive Web Scraping. In ICDSMLA 2019. Amit Kumar, Marcin Paprzycki, and Vinit Kumar Gunjan, eds. Pp. 1245–1256. Lecture Notes in Electrical Engineering. Singapore : Springer.

Mélice, Anne
2009 Un concept lévi-straussien déconstruit : le « bricolage ». Les Temps Modernes 656(5). Paris : Gallimard : 83–98.

Millstein, Frank
2020 Natural Language Processing With Python : Natural Language Processing Using NLTK. Frank Millstein.

M'zoughi, Meriem
2021 Bricolage médical. Savoirs et pratiques des oncologues cambodgiens. Moussons. Recherche en sciences humaines sur l'Asie du Sud-Est(38). Presses Universitaires de Provence : 137–165.

Nikolenko, Sergey I., Sergei Koltcov, and Olessia Koltsova
2017 Topic Modelling for Qualitative Studies. Journal of Information Science 43(1). SAGE Publications Ltd : 88–102.

Obenshain, Mary K.
2004 Application of Data Mining Techniques to Healthcare Data. Infection Control & Hospital Epidemiology 25(8). Cambridge University Press : 690–695.

Odin, Florence, and Christian Thuderoz
2010 Des mondes bricolés : arts et sciences à l'épreuve de la notion de bricolage. PPUR Presses polytechniques.

Osman, Abdullahi Sidow
2019 Data Mining Techniques : Review. International Journal of Data Science Research 2(1) : 1–5.

Palod, Priyank, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal
2019 Misleading Metadata Detection on YouTube. In Advances in Information Retrieval. Leif Azzopardi, Benno Stein, Norbert Fuhr, et al., eds. Pp. 140–147. Lecture Notes in Computer Science. Cham : Springer International Publishing.

Paredes-Valverde, Mario Andrés, Giner Alor-Hernández, Alejandro Rodr'guez-González, and Gandhi Hernández-Chan
2012 Developing Social Networks Mashups : An Overview of REST-Based APIs. Procedia Technology 3. The 2012 Iberoamerican Conference on Electronics Engineering and

Computer Science : 205–213.

Pujari, Arun K.
2001 Data Mining Techniques. Universities Press.

Racine, Valerie
2014 “Evolution and Tinkering” (1977), by Francois Jacob. https://hpsrepository.asu.edu/handle/10776/8227, accessed November 15, 2022.

Rangaswamy, Shanta, Shubham Ghosh, Srishti Jha, and Soodamani Ramalingam
2016 Metadata Extraction and Classification of YouTube Videos Using Sentiment Analysis. In 2016 IEEE International Carnahan Conference on Security Technology (ICCST) Pp. 1–2.

Rénier, Louis
2022 « Pourquoi je protège mon blé ». La communauté des agri-youtubeurs et ses publications sur les pesticides. Études rurales 209(1). Paris : Éditions de l’EHESS : 106–127.

Rénier, Louis, Aurélie Cardona, Frédéric Goulet, and Guillaume Ollivier
2022 La proximité à distance. Reseaux 231(1) : 225–257.

Research, Daniel Ellis
2022 Language Translation Using Python. Medium. https://towardsdatascience.com/language-translation-using-python-bd8020772ccc, accessed November 9, 2022.

Rogers, Matt
2015 Contextualizing Theories and Practices of Bricolage Research. The Qualitative Report. https://nsuworks.nova.edu/tqr/vol17/iss48/3/, accessed March 15, 2022.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger
2002 Multiword Expressions : A Pain in the Neck for NLP. In Computational Linguistics and Intelligent Text Processing. Alexander Gelbukh, ed. Pp. 1–15. Lecture Notes in Computer Science. Berlin, Heidelberg : Springer.

Sboui, Oumayma, Souha Kefi, Michel J. F. Dubois, and Davide Rizzo
2022 Bricolage of Agri-Equipment. In . https://hal.science/hal-03777855, accessed February 9, 2023.

Senyard, Julienne, Ted Baker, and Per Davidsson
2009 Entrepreneurial Bricolage : Towards Systematic Empirical Testing. Frontiers of Enrepreneurship Research 29(5) : 16.

Singrodia, Vidhi, Anirban Mitra, and Subrata Paul
2019 A Review on Web Scrapping and Its Applications. In 2019 International Conference on Computer Communication and Informatics (ICCCI) Pp. 1–6.

Srinivasa-Desikan, Bhargav
2018 Natural Language Processing and Computational Linguistics : A Practical Guide to Text Analysis with Python, Gensim, SpaCy, and Keras. Packt Publishing Ltd.

Stinchfield, Bryan T., Reed E. Nelson, and Matthew S. Wood
2013 Learning from Levi–Strauss' Legacy : Art, Craft, Engineering, Bricolage, and Brokerage in Entrepreneurship. Entrepreneurship Theory and Practice 37(4). SAGE Publications Inc : 889–921.

Trouvé, Alain
2020 La lecture, entre bricolage et déconstruction : autour de "La Potière jalouse" (Lévi-Strauss). Éditions et Presses universitaires de Reims. https://hal.univ-reims.fr/hal-02967387, accessed November 15, 2022.

Vasiliev, Yuli
2020 Natural Language Processing with Python and SpaCy : A Practical Introduction. No Starch Press.

Wang, Meng, and Fanghui Hu
2021 The Application of NLTK Library for Python Natural Language Processing in Corpus Research. Theory and Practice in Language Studies 11(9) : 1041–1049.

Xie, Qing, Xinyuan Zhang, Ying Ding, and Min Song
2020 Monolingual and Multilingual Topic Analysis Using LDA and BERT Embeddings. Journal of Informetrics 14(3) : 101055.

Yogish, Deepa, T. N. Manjunath, and Ravindra S. Hegadi
2019 Review on Natural Language Processing Trends and Techniques Using NLTK. In Recent Trends in Image Processing and Pattern Recognition. K. C. Santosh and Ravindra S. Hegadi, eds. Pp. 589–606. Communications in Computer and Information Science. Singapore : Springer.

Zaki, Mohammed J., and Limsoon Wong
2004 Data Mining Techniques. In Selected Topics in Post-Genome Knowledge Discovery Pp. 125–163. Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, Volume 3. CO-PUBLISHED WITH SINGAPORE UNIVERSITY PRESS. https://www.worldscientific.com/doi/abs/10.1142/9789812794840_0004, accessed November 8, 2022.

Zeni, Mattia, Daniele Miorandi, and Francesco De Pellegrini
2013 YOUStatAnalyzer : A Tool for Analysing the Dynamics of YouTube Content Popularity. In Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools Pp. 286–289. ValueTools '13. Brussels, BEL : ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). https://doi.org/10.4108/icst.valuetools.2013.254391, accessed November 9, 2022.

Zhao, Bo
2017 Web Scraping. In Encyclopedia of Big Data. C Schintler and C McNeely, eds. Pp. 1–3. Cham : Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_483-1.