Twitter comme corpus numérique d'analyse des représentations territoriales
Application au Parc national des Calanques de Marseille-Cassis-La Ciotat

AUTEUR.E.S
Siqi FAN, Philippe DEBOUDT, Amel FRAISSE, Éric KERGOSIEN

RÉSUMÉ
À partir du terrain constitué par le Parc national des Calanques, cette communication présente les objectifs, la méthodologie et les premiers résultats d'un projet de recherche interdisciplinaire (géographie, sciences de l'information, informatique) soutenu par le LabEx DRIIHM-CNRS et OHM Littoral méditerranéen. La méthodologie semi-automatisée que nous présentons vise à identifier et analyser les thématiques mentionnées et les acteurs qui s'expriment sur le territoire d'études à partir de Twitter.

MOTS CLÉS
Twitter, représentation territoriale, fouille de textes, participation, gouvernance, Parc national des Calanques

ABSTRACT
This paper presents the objectives, methodology and initial results of an interdisciplinary research project (geography, information and communication sciences) based on the site of the Calanques National Park. This project is founded by the LabEx DRIIHM-CNRS and the OHM Littoral méditerranéen. To this end, we present a semi-automatic methodology to identify and analyse descriptors related to the territory of the Calanques National Park from Twitter social network.

KEYWORDS
Twitter, territorial sciences, text mining, participation, governance, Calanques National Park

INTRODUCTION
Dans cette communication, nous présentons l'élaboration d'une méthodologie semi-automatisée visant à identifier et analyser les descripteurs permettant d'identifier et de comprendre les enjeux du territoire du Parc national des Calanques à partir du réseau social Twitter. Nous faisons dans ce sens l'hypothèse que les réseaux sociaux, et notamment Twitter, sont un moyen de communication permettant à des acteurs, peu visibles dans les médias, de s'exprimer sur des thématiques liées à l'aménagement du territoire, à l'environnement, aux politiques publiques, etc. La notion de territoire fait référence à différents concepts tels que les informations spatiales et temporelles, les acteurs, les opinions, l'histoire, la politique, etc. Dans le cadre de ces travaux, nous nous focalisons sur la détection d'entités nommées (EN) de type acteurs, lieux (que l'on nomme entités spatiales), temporel et thématique. Plus précisément, nous proposons une approche interdisciplinaire mobilisant l'expertise thématique issue de la géographie à une approche de fouille de textes issue des sciences de l'information et de la communication pour extraire les descripteurs territoriaux. Nous proposons des premiers éléments de réponse aux questions suivantes : Quels sont les acteurs qui s'expriment sur les lieux/sujets en lien avec le Parc national des Calanques ? Quelles

sont les relations entre ces acteurs ? Quelles sont les évolutions observées selon différentes temporalités ?

TRAVAUX CONNEXES

De 2008 à 2011, des recherches interdisciplinaires en sciences sociales (sociologie, géographie, aménagement-urbanisme) ont analysé le processus de construction territoriale du Parc national des Calanques de Marseille-Cassis-La Ciotat et la concertation mise en œuvre pour réaliser la charte du parc national (Deldrève & Deboudt, 2012). Le processus de création a débuté en 2007. Nos résultats de recherche démontrent la difficulté d'articuler les enjeux globaux et locaux dans la construction d'un tel projet de territoire qui a principalement bénéficié aux usagers traditionnels et s'est accompagnée d'une exclusion des enjeux urbains et maritimes. Nous avons souhaité réinterroger ces résultats (principalement obtenus à partir de méthodologies combinant des enquêtes par entretiens semi-directifs et de l'observation participante) en mobilisant des corpus de données numériques issues des réseaux sociaux (Twitter). Depuis la création du Parc national en 2012, les acteurs se sont-ils mobilisés ou exprimés sur les réseaux sociaux à son propos ? Les événements récents associés aux rejets de résidus en mer (boues rouges) dans le périmètre du parc national, par un site industriel de production d'alumine calcinée (ALTEO) à Gardanne, montrent la difficulté d'articuler dans un projet de territoire des enjeux globaux avec des enjeux locaux, et sont susceptibles de provoquer des prises de paroles sur les réseaux sociaux. Les discours, projets portés dans ces réseaux alimentent, rejoignent-ils ou s'opposent-ils aux projets inscrits dans les agendas politiques ou développés dans l'espace public physique ?

Pour répondre à ces questions, il est important de pouvoir identifier et valider les descripteurs permettant de décrire le territoire d'études, à savoir les entités nommées de type acteurs, lieux, temporalités, et thématiques.

1. EXTRACTION DES ENTITÉS NOMMÉES

Les entités nommées (EN) ont été définies comme des noms de personnes, lieux et organisations lors des campagnes d'évaluations américaines appelées MUC (*Message Understanding Conferences*) et organisées dans les années 90. Un premier défi consiste à reconnaître dans les tweets les entités nommées (EN) de type lieu, organisation et date. De nombreuses méthodes permettent de les reconnaître à partir de textes, notamment les approches statistiques qui étudient les termes co-occurrents par analyse de leur distribution dans un corpus ou par des mesures calculant la probabilité d'occurrence d'un ensemble de termes. Ces méthodes ne permettent pas toujours de qualifier des termes comme étant des EN, notamment les EN de type entités spatiales (ES) ou acteurs. Des méthodes de fouille de données fondées sur l'extraction de motifs permettent de déterminer des règles (appelées règles de transduction) afin de repérer les EN, qui utilisent des informations syntaxiques propres aux phrases. Des approches récentes s'appuient sur le web pour établir des liens entre des entités et leur type (ou catégorie). Globalement, les relations peuvent être identifiées par des calculs de similarité entre des contextes syntaxiques par prédiction à l'aide de réseaux bayésiens, par des techniques de fouille de textes ou encore par inférence de connaissances à l'aide d'algorithmes d'apprentissage. Ces méthodes sont efficaces, mais elles ne sont pas adaptées aux particularités des corpus de tweets, et notamment au langage utilisé contenant des abréviations, des fautes, et souvent peu de structure syntaxique. Enfin, pour la reconnaissance des classes d'EN, de nombreuses approches s'appuient sur des méthodes d'apprentissage supervisé, comme les machines à vecteurs de support (support vector machine ou SVM)

ou encore les champs aléatoires conditionnels (conditional random fields ou CRF) qui sont souvent utilisées dans le challenge *Conference on Natural Language Learning* (CoNLL). Les algorithmes exploitent des données expertisées/étiquetées ainsi que divers descripteurs, comme par exemple les positions des termes, les étiquettes grammaticales, les informations lexicales (eg. majuscules/minuscules), les affixes, l'ensemble des mots dans une fenêtre autour du candidat, etc. Bien qu'intéressantes, ce type d'approche nécessite un travail manuel d'étiquetage important que nous ne pouvons appliquer dans le cadre de ces travaux.

Plusieurs travaux de recherche utilisent Twitter comme corpus pour construire des ressources linguistiques et extraire des connaissances pertinentes. Par exemple, Read (2005), Pak et Paroubek (2010) ont utilisé les émoticônes comme marqueur de polarité pour distinguer les textes positifs et négatifs depuis les tweets. Mohammad (2012), Qadira et Riloffe (2013), Fraisse et Paroubek (2014) ont utilisé une liste de mot-dièses ou hashtags (eg. #sad, #happy, #angry, #fear, #anxious, #disappointed, #unhappy, etc.) pour collecter des corpus émotionnels et construire de façon automatique des lexiques affectifs. Les lexiques ont été ensuite utilisés dans des tâches de détection automatique des émotions. Zenasni *et al.* (2016) propose une approche de fouille de textes pour identifier et extraire les entités nommées dans les corpus de messages courts (SMS et tweets) en prenant en compte l'évolution du langage caractérisant ce type de textes non standards.

Dans cette recherche, nous proposons une méthode combinant une approche statistique à une approche de fouille de textes (Pak *et al.*, 2014 ; Zenasni *et al.*, 2016).

2. UNE MÉTHODE POUR LA CONSTRUCTION D'UNE REPRÉSENTATION TERRITORIALE

Pour la collecte du corpus, deux périodes temporelles ont été choisies :
– La période 2007-2011 correspondant à celle du processus de création du Parc national des Calanques et de l'organisation du processus de concertation pour élaborer la charte du parc national. C'est durant cette période que se sont manifestés des groupes d'acteurs favorables ou opposés à la création d'un parc national dans le territoire des Calanques.
– La période 2012-2017 intégrant l'année de création du Parc national (2012) jusqu'aux années récentes marquées par plusieurs conflits d'usages et notamment celui fortement médiatisé provoqué par des rejets de boues rouges issues de la production d'alumine par l'usine ALTEO de Gardanne dans les espaces du cœur maritime du Parc national.

Avant de lancer la collecte automatique du corpus de tweets, nous avons identifié manuellement, avec l'expertise de géographes spécialistes du territoire d'études, une liste d'acteurs qui ont participé au processus de création du parc. Il s'agit essentiellement d'associations, de conseils de quartiers, de personnalités politiques et de citoyens. Nous avons aussi identifié une liste de mots clés que nous avons utilisés sous forme de hashtags pour collecter les tweets, comme par exemple #PNCalanques, #ParcNationalDesCalanques, #BouesRouges, etc. (tabl. 1). Nous avons utilisé l'API Search[1] de Twitter pour collecter et filtrer les messages, qui permet de spécifier la langue de messages et une requête de recherche par mot clé. Ainsi, pour chaque hashtag *h* du tableau 1, nous collectons un certain nombre de tweets qui contiennent le hashtag *h*. Au total, sur la période 2007-2011 (avant la création du parc national) nous avons collecté 5 000 tweets et pour la 2e période, 2 900 tweets.

1 dev.twitter.com/docs/api/1/get/search

Avec l'aide des experts du domaine, nous avons analysé dans un premier temps le corpus collecté afin de vérifier la pertinence des hashtags utilisés pour la collecte des tweets, ce qui nous a permis d'en modifier et d'en rajouter avant de relancer le processus de collecte.

Tableau 1. Extrait de la liste de hashtags utilisés pour la construction de corpus de tweets

Hashtags	Description
#ParcNationalCalanques	Parc national des Calanques
#PNCal	Parc national des Calanques
#BouesRouges	boues rouges
#Marseille	Marseille
#CreationPNCalanques	création du Parc national des Calanques

Tableau 2. Exemples de tweets extraits du corpus collecté

Exemples de tweets collectés
@cestrosi @MetropoleNCA #stephaneBouillon préfet #paca qui autorise la pollution dans la méditerranée de l'entreprise #altéo #bouesrouge
#ComitéSantéLittoral Sud (Marseille): Bulletin n°1 http://comitesantelittoralsud.blogspot.com/2014/01/bulletin-n1.html?spref=tw …
#marseille: Parc national des Calanques : les élus face à l'enquête publique http://bit.ly/sxqeOx
la création du #PNCalanques aura permis de préserver l'espèce des avocats et juristes en tout genre ! http://www.marsactu.fr/environnement/parc-des-calanques-un-recours-depose-contre-lelection-de-danielle-milon-30167.html …

En se basant sur une approche statistique, notre méthode consiste à extraire, à partir du corpus collecté, et pour chaque hashtag *h*, l'ensemble de mots qui lui est associé. Ces mots peuvent être des noms d'acteurs, des mots simples ou des noms de lieux. En effet, nous considérons que si un mot *m* est fortement corrélé à un hashtag *h* de notre liste alors ce mot est pertinent pour notre analyse de relations territoriales. Afin de mesurer cette association, nous nous sommes basés sur l'information mutuelle introduite par Fano (1961) qui, pour chaque couple de variables aléatoires (X, Y), mesure leur degré de dépendance au sens probabiliste. L'information mutuelle est donnée par la formule suivante :

$$IM(X,Y) = \log_2\left(\frac{P(X,Y)}{P(X).P(Y)}\right)$$

Ainsi, dans notre cas, il s'agit de mesurer le degré de dépendance entre un hashtag *h* et un mot *m*.

$$IM(h,m) = \log_2\left(\frac{freq(h,m)}{freq(h).freq(m)}\right)$$

freq(h,m) est le rapport entre le nombre de tweets contenant le mot *m* et le hashtag *h* ($|T_{h,m}|$) et le nombre total de tweets ($|T|$).

$$freq(h,m) = \frac{|T_{h,m}|}{|T|}$$

freq(h) est le rapport entre le nombre total de tweets contenant le hashtag *h* ($|T_h|$) et le nombre total de tweets.

$$freq(h) = \frac{|T_h|}{|T|}$$

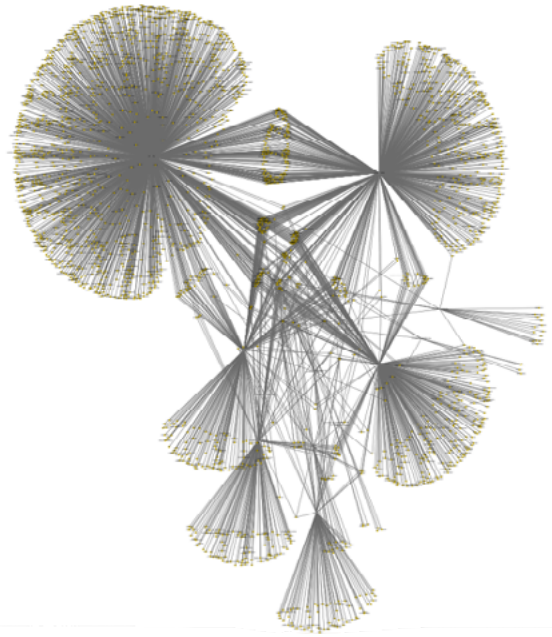
freq(h) est le rapport entre le nombre total de tweets contenant le mot *m* ($|T_m|$) et le nombre total de tweets.

$$freq(m) = \frac{|T_m|}{|T|}$$

3. EXPÉRIMENTATION ET RÉSULTATS

Les messages Twitter peuvent contenir des URL, des mentions utilisateurs (les acteurs dans notre cas, par exemple, @MairiedeMarseille), des retweets, etc. Ainsi, avant de procéder à l'extraction des hashtags, des noms d'acteurs et des lieux cités dans les tweets, nous avons tout d'abord effectué certaines opérations de prétraitements automatiques : (1) suppression des liens URL et des retweets, (2) segmentation et (3) suppressions des mots outils. Ensuite, nous avons calculé pour chaque mot *m* du corpus sa corrélation avec le hashtag *h* qui a été utilisé dans la requête. En fonction de leurs degrés d'association, les mots sont ensuite ordonnés par ordre croissant : du plus pertinent au moins pertinent.

Figure 1. Visualisation des relations entre les termes liés à la thématique « environnement »

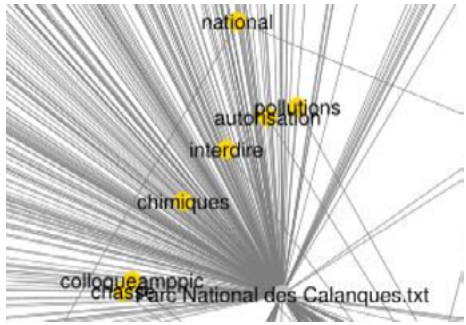


Afin de mieux visualiser les résultats obtenus, nous avons fait appel à un expert du domaine pour fixer un seuil en dessous duquel nous considérons qu'un mot *m* n'est pas pertinent dans notre analyse.

Nous avons procédé ensuite à la visualisation de ces résultats *via* des graphes : les nœuds représentent les mots, les hashtags et les noms d'acteurs dans le corpus, et les arcs sont soit :
– Les relations entre les mots (mot clé ou hashtag) : cette relation permet de répondre à la question « Quelles sont les thématiques et les sujets abordés dans les tweets ? », par exemple les relations entre termes autour de la thématique « Pollution » (fig. 1) et plus précisément les relations entre « Parc national des calanques » et « pollution » (fig. 2).
– Les relations entre un hashtag et un acteur : cette relation permet de répondre à la question « Qui parle de quoi ? ».

– Les relations entre deux acteurs : cette relation permet de répondre à la question « Qui parle à qui ? ».

Figure 2. Visualisation des relations entre les hashtags et les mots clés du corpus



CONCLUSION ET PERSPECTIVES

Cette communication a permis de livrer le processus de construction de la méthodologie pour créer le corpus de données numériques, identifier les connaissances pertinentes et nos premiers résultats concernant les thèmes, les acteurs, leurs interactions à propos du Parc national des calanques dans le réseau social Twitter. La mobilisation des données numériques de Twitter permet de constituer un nouveau corpus de données numériques contemporain de la période de création du Parc national des Calanques selon une démarche rétrospective, et aussi de fournir des données numériques associées aux premières années de fonctionnement du parc national. L'interprétation fine des données récoltées, en cours de réalisation par les membres du projet, devrait permettre de compléter les analyses déjà réalisées et de produire de nouveaux résultats concernant : les interactions entre les acteurs (Berthelot *et al.*, 2016), les thématiques et les lieux, à propos du fonctionnement du parc national.

RÉFÉRENCES

Berthelot M.-A., Severo M., Kergosien É., 2016, « Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile du Nord-Pas-de-Calais », *Proceedings du 3e colloque international du CIST*, Paris, CIST, p. 66-72 [en ligne : hal.archives-ouvertes.fr/hal-01353660].

Deldrève V., Deboudt P. (coord.), 2012, *Le parc national des Calanques. Construction territoriale, concertation, usages*, Paris, Quae.

Fano R., 1961, *Transmission of Information: A Statistical Theory of Communications*, Cambridge, MA, MIT Press.

Fraisse A., Paroubek P., 2014, « Twitter as a Comparable Corpus to build Multilingual Affective Lexicons », *Proceedings of the 7th International Workshop on Building and Using Comparable Corpora* (BUCC 2014), Reykjavik, Islande, p. 17-21.

Mohammads M., 2012, « Emotional tweets », *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, p. 246-255.

Pak A., Paroubek P., 2010, « Construction d'un lexique affectif pour le français à partir de twitter », *Proceedings of TALN (Traitement Automatique des Langues Naturelles)*, Montréal, Canada, p. 6.

Pak A., Paroubek P., Fraisse A., Francopoulo G., 2014, « Normalization of Term Weighting Scheme for Sentiment Analysis », *in* Z. Vetulani et J. Mariani J. (dir.), *Human Language technology Challenges for Computer Science and Linguistics*, Springer, vol. 8387.

Qadir A., Riloff E., 2013, « Bootstrapped learning of emotion hashtags hashtags4you », *4th Workshop on Computational Approaches to Subjectivity Sentiment and Social Media Analysis*, Atlanta, p. 10.
Read J., 2005, « Using emoticons to reduce dependency in machine learning techniques for sentiment classification », *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, p. 43-48.
Zenasni S., Kergosien É., Roche M., Teisseire M., 2016, « Extracting new Spatial Entities and Relations from Short Messages », *8th International ACM Conference on Management of Digital EcoSystems* (MEDES'2016), Hendaye, p. 8.

LES AUTEUR.E.S

Siqi Fan
Université d'Orléans
fan.siqi18@gmail.com

Philippe Deboudt
Université de Lille – TVES
philippe.deboudt@univ-lille1.fr

Amel Fraisse
Université de Lille – GERiiCO
amel.fraisse@univ-lille3.fr

Éric Kergosien
Université de Lille – GERiiCO
eric.kergosien@univ-lille3.fr