



**HAL**  
open science

# Echoes des publications scientifiques en SHS sur les réseaux sociaux. Le cas des contenus d'Open Édition sur Twitter.

Lucie Loubère, Fidelia Ibekwe

## ► To cite this version:

Lucie Loubère, Fidelia Ibekwe. Echoes des publications scientifiques en SHS sur les réseaux sociaux. Le cas des contenus d'Open Édition sur Twitter.. 12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?, Oct 2019, Montpellier, France. hal-02307664

**HAL Id: hal-02307664**

**<https://hal.science/hal-02307664v1>**

Submitted on 7 Oct 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Echoes des publications scientifiques en SHS sur les réseaux sociaux. Le cas des contenus d'Open Édition sur Twitter.

## **Lucie Loubère**

Docteure en Sciences de l'information et de la communication  
Post-doctorante sur le projet Appropriations of Open Knowledge  
Aix Marseille Univ, Université de Toulon, IMSIC, Marseille, France.

[lucie.loubere@univ-amu.fr](mailto:lucie.loubere@univ-amu.fr)

## **Fidelia Ibekwe**

Professeure des universités en Sciences de l'information et de la communication  
Aix Marseille Univ, Université de Toulon, IMSIC, Marseille, France

[fidelia.ibekwe@univ-amu.fr](mailto:fidelia.ibekwe@univ-amu.fr)

## Résumé

Les réseaux sociaux en se diffusant sur l'intégralité de la société sont également entrés dans le monde de la recherche. Ces outils accélèrent la circulation de l'information, et pourraient atteindre une audience différente du circuit universitaire. Parallèlement les plateformes de savoir ouvert se développent et rendent accessible à tout le monde le savoir scientifique. Notre étude se focalise sur l'étude des tweets émis entre 2013 et 2017 pointant vers un contenu d'OpenEdition. Nous avons analysé les réseaux de retweets ainsi que les contenus textuels des tweets par étude lexicométrique. Les résultats tendent à montrer une conservation des pratiques institutionnelles avec un cloisonnement linguistique et disciplinaire tant dans les pratiques de retweets que dans les contenus textuels.

## Mots clés

Science ouverte, twitter, cluster, lexicométrie, altmetrics

## Title

Echoes of scientific publications in SHS on social networks. The case of Open Edition content on Twitter.

## Abstract

Social networks, spreading throughout society, have also entered the world of research. These tools accelerate the flow of information, and could reach a different audience within the university circuit. At the same time, open knowledge platforms are developing and making scientific knowledge accessible to everyone. Our study focuses on the study of tweets issued between 2013 and 2017 pointing to OpenEdition content. We analyzed the retweets networks as well as the textual content of the tweets by lexicometric study. The results tend to show a conservation of institutional practices with linguistic and disciplinary compartmentalization in both retweets practices and textual content.

## Keywords

open science, twitter, lexicometry, altmetrics

## INTRODUCTION

Les réseaux sociaux numériques sont devenus un canal accepté de diffusion de la science. De nombreux chercheurs possèdent un compte Twitter (Costas, van Honk, & Franssen, 2017) et l'utilisent pour diffuser les résultats des recherches scientifiques. Ces contributions sont parfois retweetées et peuvent circuler en dehors de la sphère académique. Ce mode de diffusion, accentué par la prolifération de veilleurs à l'affût de nouvelles publications, interroge les modalités d'organisation des connaissances. Jusqu'à l'arrivée des moteurs de recherche et des plateformes de web 2.0, l'organisation des connaissances et la recherche documentaire reposaient sur les systèmes de classification construits par des spécialistes en Sciences de l'information et des bibliothèques (*Library & Information Science*) dans une approche hiérarchique (*top-down*), souvent encyclopédique et normative (privilégiant une seule vision de l'organisation des connaissances). Ces systèmes doivent désormais composer avec les plateformes de web 2.0 qui ont installé un fonctionnement horizontal à l'organisation des connaissances par l'intégration des contributions du grand public, entraînant ainsi l'effacement des frontières entre professionnels et amateurs (Ibekwe-SanJuan & Geoffrey, 2017, p. 196).

Parallèlement à la virtualisation des échanges, l'ouverture des contenus scientifiques en mode « open access » peut être perçue comme une opportunité pour les chercheurs et l'ensemble de la société, car elle permet d'accélérer la diffusion des résultats de la science en société et par là, sa percolation et traduction en savoir et en innovation.

Les méthodes et outils bibliométriques tentent de prendre en compte la diversité des supports de publications en ajoutant de nouveaux indicateurs (Hicks, Wouters, Waltman, Rijcke, & Rafols, 2015; O'Neill, 2016; Robinson-Garcia et al., 2019). Cependant ces nouvelles métriques (dites « altmetrics ») si elles donnent un aperçu quantitatif de l'écho que reçoit une publication scientifique dans le cyberspace, ne nous renseignent pas sur les modalités de ses réutilisations.

Cette étude se situe à l'interface des analyses de réseaux sociaux numériques et des recherches en bibliométrie. Nous souhaitons interroger les contenus présents sur twitter et qui font mention de la science ouverte afin d'étudier quels sont les sujets abordés mais également par qui. Les contenus scientifiques, une fois mis à disposition et partagés sur ces plateformes, sortent-ils réellement de la sphère académique, ou reproduisent-ils les mêmes structures institutionnelles (communautés universitaire, convergence disciplinaire...)?

Notre communication s'inscrit dans le cadre du projet « Appropriation du savoir ouvert » financé par l'initiative d'Excellence (IDEX) de l'Université de Marseille (Amidex). L'objectif de ce consortium pluridisciplinaire est d'étudier les contextes de réappropriation des contenus d'OpenEdition (OE) dans la sphère publique. Le travail présenté ici porte spécifiquement sur les tweets citant un contenu scientifique publié sur une des plateformes de OE.

## 1 - METHODOLOGIE

Notre travail s'appuie sur les méthodes quantitatives et plus précisément lexicométriques. Ce courant d'analyse reposant sur la mesure du lexique des textes et le calcul statistiques des cooccurrences est une des expressions de notre démarche exploratoire (Lejeune, 2010). Habituellement employé sur des corpus construits sur un périmètre thématique délimité, ces méthodes permettent alors de faire émerger les prises de position et les controverses publiques. Dans notre étude, le corpus n'est pas délimité par un sujet spécifique, mais par la présence d'une trace de citation des contenus d'OpenEdition. Par conséquent, notre corpus peut porter sur autant de thématiques qu'il en existe dans les contenus d'OpenEdition.

Cette hétérogénéité pluridisciplinaire justifie le recours aux outils basés sur les cooccurrences. Ces dernières peuvent se définir par la « *présence simultanée de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d'un même contexte linguistique [ici le tweet]* » (Mayaffre, 2008, p. 98). La prise en compte non plus des mots isolés, mais des mots régulièrement employés avec d'autres mots permet par cette contextualisation de désambiguïser le lexique. Ainsi le mot « droit » pourra se retrouver dans une classe de discours contenant également les mots « légal », « justice » (lexique juridique) mais également dans le discours politique avec les mots « gauche », « gouvernement », « assemblée ».

## 1.1 Recueil du corpus

Le recueil des tweets couvre la période du 01/01/2013 au 31/12/2017, nous avons utilisé une extension de Google sheet développée par Martin Hawksey [1] qui permet d'archiver les résultats de l'API Twitter. Cet outil recueille « au fil de l'eau » les tweets correspondant à une requête renseignée. Nous avons effectué une requête ciblant la présence d'un lien contenant la racine d'une de ces plateformes :

- « [hypothese.org](http://hypothese.org) » pour les carnets de recherche,
- « [books.org](http://books.org) » pour les livres,
- « [journals.org](http://journals.org) » pour les revues [3],
- « [calenda.org](http://calenda.org) » pour les annonces d'événements scientifiques.

Cette collecte a abouti à 830 085 tweets émis dont 375 994 sont des tweets originaux et 454 091 des retweets. Nous avons ensuite formaté ces tweets afin d'effectuer les analyses voulues (voir ci-après). Pour chaque tweet, nous avons consigné en métadonnée le nom du compte qui l'a émis, la date de publication et la plateforme d'OE citée.

## 1.2 Traitement des données

À partir des données récoltées par l'api, nous avons procédé au traitement de ces dernières en deux temps que nous détaillerons ici. La première étape consiste en l'extrapolation de clusters sur la base de « qui retweete qui ». Le second temps est consacré à une étude des contenus textuels des tweets selon une classification hiérarchique descendante (CHD) de type reinert (Reinert, 1983).

### 1.2.1 Calcul des clusters de retweets

Cette méthode souvent étudiée dans le courant des Digital Methods est basée sur les interactions entre les personnes sur un même réseau social. Ce travail repose sur des données quantitatives, les traces numériques laissées par les usagers lorsqu'ils retweetent une tierce personne, ou plus précisément un tiers compte. Nous avons donc sur l'ensemble des 830 085 tweets identifiés l'origine et destination des retweets et leurs fréquences pour constituer un graphe synthétisant tous ces liens. L'algorithme utilisé ici est la méthode de Louvain (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008) implémenté dans le logiciel Gephi (Bastian, Heymann, & Jacomy, 2009). Le principe de son fonctionnement est constitué de deux phases : la première consiste en la création d'un cluster par sommet, puis chaque sommet intégrera la communauté de son voisin.

Si cette structuration permet un gain de modularité, elle est conservée, si non, elle est

annulée. Ce processus itératif continue jusqu'à parvenir à une modularité maximale (aucun changement ne l'améliore). La seconde phase, reproduit le même schéma en considérant les communautés calculées précédemment comme des sommets (singletons) à part entière, les arêtes reliant ces derniers sont calculées par la somme des poids des liens entre les individus constituant les deux clusters. Ces groupements de comptes nous permettent donc d'assigner aux comptes auteurs des tweets et par conséquent aux tweets eux-mêmes l'identification d'une proximité de pratique de publication sous la forme de clusters. L'appartenance d'un compte à un cluster est consignée pour intégrer les métadonnées dans l'analyse du contenu textuel.

### 1.2.2 Classification hiérarchique descendante.

La classification hiérarchique descendante (CHD) employée ici est celle implémentée dans le logiciel Iramuteq (Ratinaud, 2009). Elle permet de mettre à jour les mondes lexicaux qui partent du principe que l'énoncé est un point de vue dépendant du sujet mais aussi de son activité et de son contexte où :

*« le vocabulaire d'un énoncé particulier [est considéré] comme une trace pertinente de ce « point de vue » il est à la fois la trace d'un lieu référentiel et d'une activité cohérente du sujet-énonciateur. Nous appelons mondes lexicaux, les traces les plus prégnantes de ces activités dans le lexique. » (Reinert, 1993, p. 11).*

Afin d'augmenter la récurrence des mots, cet outil va lemmatiser le corpus en réduisant les verbes à l'infinitif, et les noms, adverbess et adjectifs au masculin singulier. Ensuite, un tableau lexical croise la présence ou absence des formes pleines (verbes, nom, adjectif, adverbess) dans les tweets. Sur ce tableau est effectuée une série de bipartitions sur la base d'une Analyse Factorielle de Correspondances (AFC). Ceci permet de regrouper les segments les plus proches sur un critère de cooccurrence lexicale. Une fois ce découpage effectué, nous obtenons les « classes de discours ». Pour chacune d'entre elles, est construite et présentée la liste des formes et des métadonnées statistiquement surreprésentées. Si nous avons ajouté en métadonnées le cluster d'appartenance de l'auteur du tweet, nous rappelons que cette variable n'est pas employée dans la structuration du tableau lexical. La structuration des classes de discours se fait donc indépendamment de ces données, ce n'est qu'a posteriori que le calcul de corrélation classe/modalité de la métadonnée est effectué.

## 2 – LES RESULTATS

### 2.1 Analyse des clusters

Le graphe présenté en illustration 1 représente les huit plus grands clusters calculés sur les comptes ayant généré au moins cinq tweets. Il est spatialisé par l'algorithme Alpha [2], les comptes visibles sont les plus retweetés par cluster.

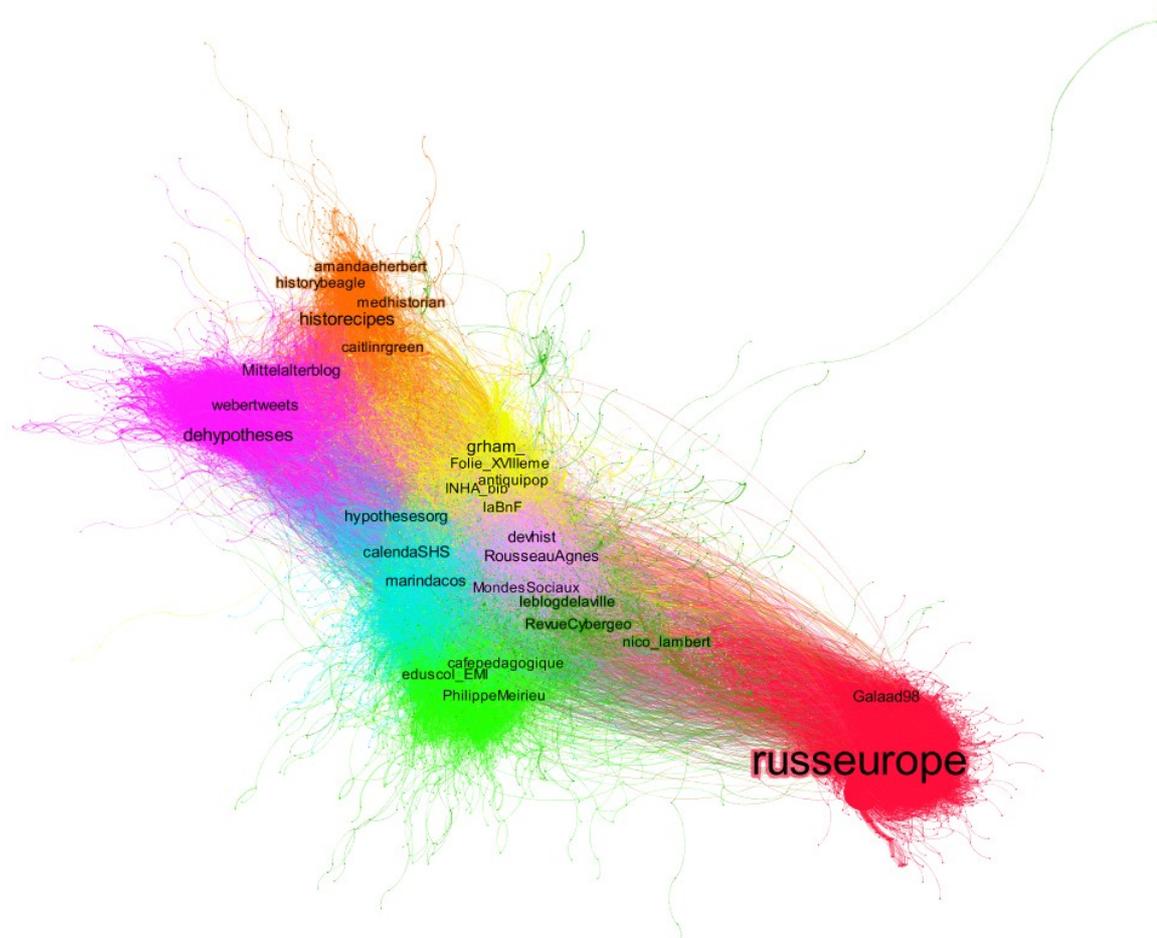


Illustration 1 : graphe de clusters de retweets

Les rapprochements de comptes que délimite cette illustration mettent en évidence plusieurs points. Une première distinction se fait sur la langue du compte. Les clusters en haut des graphes colorés en rose et orange sont respectivement constitués de comptes germanophones et anglophones. Le partage de contenus semble donc respecter la barrière de la langue. Nous noterons que l'élargissement de cette analyse à l'ensemble des comptes (dont ceux ayant émis peu de tweets) fait apparaître des regroupements hispaniques, italiens et encore d'autres langues. Le second critère saillant dans la répartition des comptes est l'appartenance disciplinaire. La plateforme d'OpenEdition regroupant l'ensemble des sciences

humaines et sociales, nous voyons ressortir des groupements basés sur des comptes aux identités disciplinaires marquées.

Ainsi le cluster vert foncé en se formant autour des comptes « revuedelaville », « cybergéo » et « nico\_lambert » est concerné par les sciences géographiques. Le deuxième cluster vert clair en dessous est quant à lui centré sur les comptes « eduscol », « philippeMeirieu », « cafe pedagogique », autant de comptes évoluant dans le domaine éducatif dont les sciences de l'éducation. Les clusters roses et jaunes regroupent également des tweets issus de plusieurs disciplines. Le premier est orienté sciences sociales avec le compte « mondessociaux », « devhist » et « rousseauAgnes » alors que le second suit de façon prononcée les comptes de chercheurs en histoire, de bibliothèques (universitaires ou non) ou d'autres organismes d'archives.

Enfin deux autres groupes se détachent : le cluster bleu orienté autour des comptes officiels d'OpenEdition (« Marindacos », « hypothese », « calenda ») recouvre les tweets émis par les membres de cette organisation. Le cluster rouge lui est centré sur le compte et carnet de recherche RusseEurope. Ce dernier groupement est particulièrement dense comme le montre l'illustration. Il se détache des autres clusters, cette spatialisation montre le peu de liens entretenus entre les comptes de ce cluster et ceux des autres.

## 2.2 Les contenus des tweets

La seconde étape de notre étude porte sur les contenus même des tweets. Pour cela, nous appliquons la CDH. Une première classification sur l'intégralité des tweets est présentée en illustration 2. Nous avons obtenu 60 classes en tout. Dans la figure ci-après, nous ne montrons que les classes contenant un minimum de 5000 tweets, soit 20 classes.

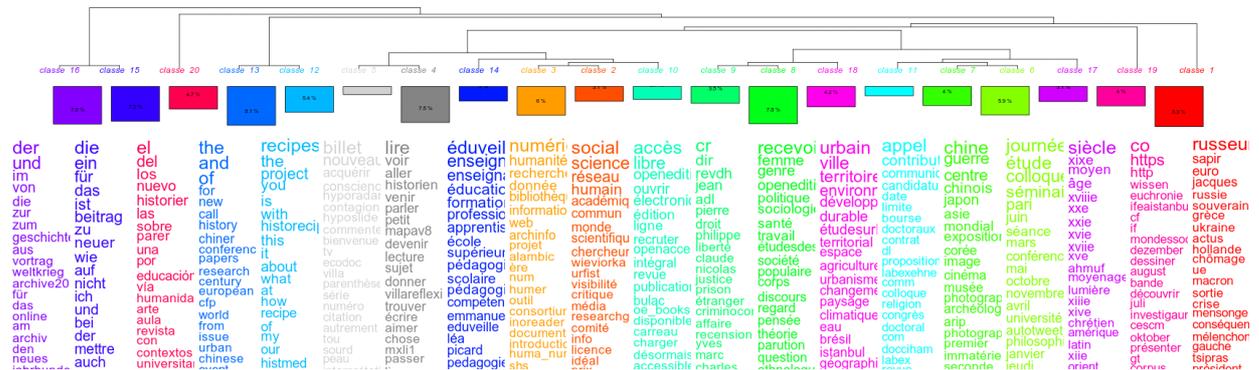


Illustration 2 : dendrogramme de la classification sur l'intégralité des tweets

La classification obtenue met en exergue, sur les premiers facteurs, la distinction des discours germanophones (classe 15 et 16 violet et bleu) sur la gauche du dendrogramme, suivis par les tweets hispanophones (classes 20 en rouge) et anglophones (classes 12 et 13 en bleu). Cette répartition ne provient pas d'une affinité thématique mais d'une limite de la CHD. En effet, la majorité des tweets émis étant en français, nous avons utilisé les dictionnaires de cette langue. Dès lors, les mots d'une autre langue n'étant pas répertoriés dans la base francophone, ils sont considérés comme des formes pleines (au même titre que des noms propres).

La récurrence des mots outils provoque dès les premières étapes de la classification une partition respectant la barrière linguistique. Dans la suite de ce travail, nous nous focaliserons sur les classes francophones, mais l'étude dans cette classification des répartitions des clusters de retweet sur les classes de discours nous montre une très forte corrélation entre les clusters vus précédemment et les classes dans une langue étrangère. Ainsi, le cluster

germanique et les classes 15 et 16 affichent une très forte corrélation [3] puisque plus de 88 % des tweets composant la classe 15 proviennent de ce cluster, ce taux s'élève à 94 % pour la classe 16 (voir Illustration 3 ci-après).

Cette tendance n'est pas retrouvée dans les classes portant un discours anglophones. L'internationalité de cette langue est retrouvée ici puisque les tests statistiques entre les classes et les clusters sont certes significatifs, mais l'étude des fréquences cumulées montre la présence de nombreuses communautés pour les tweets anglophones regroupés en classes 12 et 13 (illustration 3). La classe 13 contient par exemple une minorité de tweets émis par les membres du cluster anglophone, elle est en fait composée de nombreux appels à communication et/ou participation relayés en anglais :

segments caractéristiques de la classe 13 :

*call for papers the paris peace conference of 1919 and the challenge of a new world order*

*call for paper art esa research network 37 on urban sociology gt conference title inequality and uncertainly.*

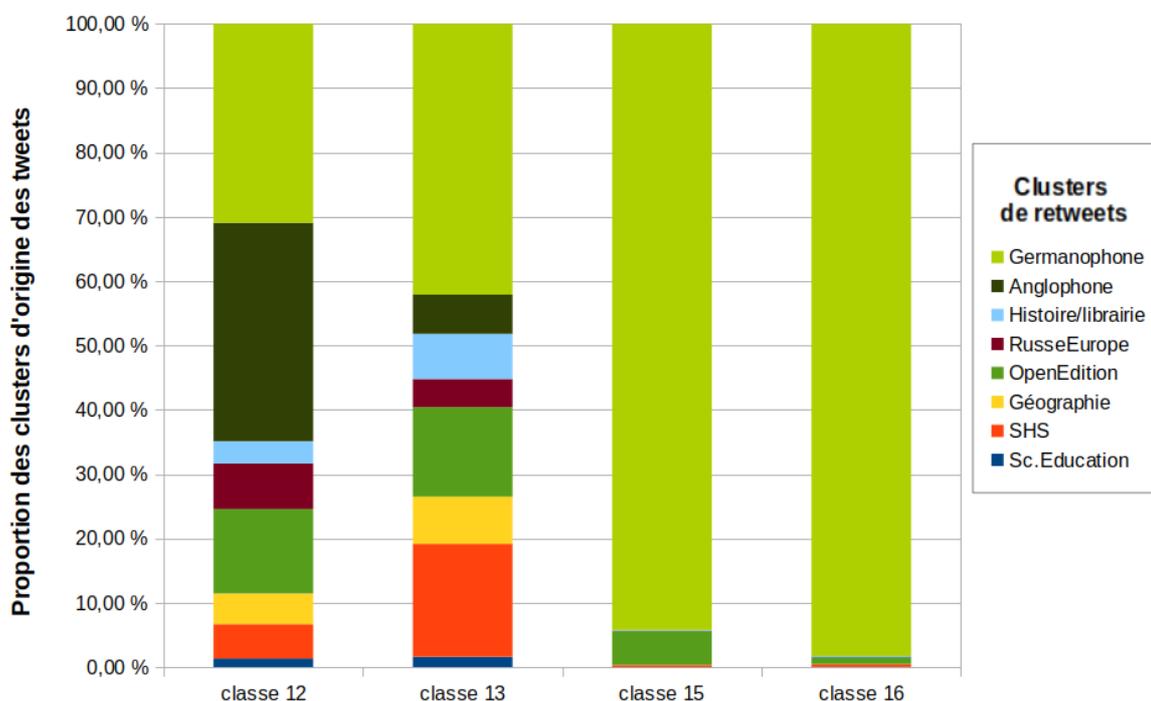


Illustration 3 : proportion des tweets par cluster d'appartenance sur les classes 12, 13, 15 et 16.

Notre objectif étant de pouvoir étudier plus précisément les différentes thématiques qui sont portées sur Twitter lors de citations des publications hébergées sur OpenEdition, nous nous focaliserons uniquement sur les classes francophones. Pour cela, nous utiliserons les classifications successives (Loubère, 2014; Ratinaud & Marchand, 2015) en exportant un sous-corpus composé uniquement des tweets contenus dans les classes concernées (soit, les classes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14, 17, 18 et 19).

Ce nouveau corpus est composé de 216 633 tweets contenant 119 261 formes répétées sur 2 771 670 occurrences. Une nouvelle CHD est produite avec 60 classes demandées et 5000 tweets minimum par classes. Le dendrogramme est présenté en illustration 4.



Extrait des tweets présents en classe 4 et 5 :

*très complet guide du droit d'auteur des chercheurs des bases de données et des archives ouvertes*

*modernisation et archives, stockage des données électroniques sae et système d'information des archives*

*enrichissement de la base patrimoine, à quand une base de données centralisée pour les archives françaises*

*2 offres d'emploi dans le domaine des humanités numériques digital humanities à l'iha digitalhumanities postdoc*

Mais au-delà des besoins matériels et humains, nous retrouvons également les nécessaires remises en questions des disciplines, de leurs enseignements et les nouveaux ponts interdisciplinaires possibles.

Extrait des tweets présents en classe 4 et 5 :

*enseigner les langues anciennes à l'ère digitale, apprendre par et pour les humanités numériques*

*humanités numériques une nouvelle discipline universitaire digital humanities à l'iha*

*les humanités numériques par le design, les usages plutôt que les outils design and digital humanites*

*humanités numériques architecture de l'information des réponses à changement de paradigme doc jpmague archinfo dh*

Si les classes précédentes portent sur l'ensemble d'acteurs de l'open science, la classe 12 est centrée sur la promotion et définition même d'OpenEdition. Dans les tweets regroupés ici sont présentes les annonces de lancement de plateformes, la promotion du nombre de documents en accès libre, mais également les enjeux du libre accès aux publications scientifiques.

Segments de textes caractéristiques de la classe 12 :

*openedition books plateforme de livres sciences humaines et sociales de la moitié d'entre eux est en libre accès*

*what's love got to do with open acces un débat sur l'accès libre aux revues en sciences sociales*

Cette classe contient de façon significative des tweets émis par les comptes présents dans le cluster d'OpenEdition mais également dans le cluster éducatif. Les tweets provenant de ces comptes sont axées sur l'éducation au média et les recommandations de lecture.

Extrait de tweets portés par le cluster éducatif et classés en classe 12

*lancement d'open édition books plateforme de livres en sciences humaines et sociales en libre accès openaccess*

*comment certains réseaux sociaux associent-ils identité numérique et eputation bonne lecture*

*l'enseignement de l'informatique pour les humanités amp les sciences sociales retour sur les journées péda*

*comment lire intelligemment une publication scientifique en shs*

*l'éducation aux médias contributions pratiques et perspectives de recherches en sciences de openeditionsays*

La classe 15 est significativement portée par les comptes présents dans le cluster éducatif. L'ouverture de la littérature scientifique impacte le monde éducatif (au sens large) sous plusieurs axes dont un apport disciplinaire aux sciences de l'éducation, mais également aux sciences de l'information et de la communication. Les contenus publiés dans OpenEdition portent pour certains sur des domaines de ces deux disciplines (éducation aux médias, didactiques...). Nous retrouvons donc les recommandations de lectures, la diffusion de nouveaux contenus par une dynamique de veille ou encore la diffusion d'appel à communication dans ces domaines.

Extrait de tweets présents dans la classe 15 :

*sur eduveille formation professionnelle des enseignants anglais écoles versus universités*

*un blog passionnant eduveille autour des recherches en éducation et formation educife*

*prolongation du délai de soumission des propositions pour le edcamp les humanités numériques pour l'éducation*

Mais le domaine éducatif ne se limite pas à un seul champ disciplinaire ou un terrain d'étude scientifique, il est bien entendu constitué de personnes concernées par la pédagogie et l'enseignement mais qui proviennent de toutes les disciplines et horizons scientifiques. L'ouverture des contenus d'OpenEdition permet ici de les proposer comme ressources pédagogiques (enseignants vers les élèves/étudiants), mais également comme contenus de formation pour le corps enseignant, en formation initiale ou déjà en poste :

Extrait de tweets présents dans la classe 15 :

*cours de formation sur l'histoire du livre et de l'imprimerie moyenage*

*ressources pour l'enseignement la formation enseignants formateur éducateurs  
chercheur léa educife*

*archivage de la séquence afrique du mooc archinfo googlealert*

*petite bibliographie sur les cours magistraux pour ceux qui s'ennuient vraiment  
trop*

*des recherches pertinentes pour les profs eduveille autour des recherches en  
éducation*

*quelles sont les recherches pertinentes pour les enseignants séminaire international  
le 31 mars*

*plateforme pour permettre la découverte de pratiques et aider à l'approche par  
compétence*

Le second groupe de discours constitué des classes 2, 8, 7, 11, 14 et 16 est centré sur les thématiques d'histoire, de géographie, de la philosophie, de l'information et communication et de la culture muséale. Nous retrouvons ici le même fonctionnement que dans la classe sur la thématique éducative. En effet, les savoirs d'histoire, géographie et philosophie composent des périmètres de disciplines scientifiques mais sont également enseignés dans le système scolaire. Il co-existera donc dans les classes de discours des tweets diffusant le contenu d'OpenEdition comme ressource pédagogique :

*est-ce que l'épreuve d'histoire et géographie en sti2d et stl intéresse quelqu'un là-  
haut*

*capex agreg parution villes et industrie géographie systèmes productifs capex agreg*

*faire cours au temps des attentats docere*

*culture visuelle éléments de cours*

Au-delà de cette spécificité, nous retrouvons les caractéristiques de l'actualité scientifique dont les appels à communication avec les classes 7, 8 et 14, qui si elles partagent la même fonction, correspondent à une spécialisation lexicale dans les disciplines (histoire, philosophie, information et communication).

Exemple de tweets de la classe 8 :

*journée d'étude interpréter l'histoire dans les arts décoratifs au XVIIIe siècle jeudi  
15 juin*

Exemple de tweets de la classe 7 :

*philosophie colloque international iris marion young 31 mai 2 juin 2017 paris*

Exemple de tweets de la classe 14 :

*appel à contribution revue études de communication les médiations des  
patrimoines vers la culture numérique*

Le dernier groupe de discours présent est centré sur les sciences sociales. Il est composé des classes 1,10, 13, 6 et 9 et présente les champs disciplinaires du droit, du langage et de la sociologie. Nous retrouvons comme sur les groupes de discours précédents des recommandations de lecture et les annonces de nouvelles parutions (classe 6 et 9), et la diffusion des nouvelles parutions, centrée sur les auteurs et/ou le titre des articles (classe 1,10 et 14).

## CONCLUSION

Cette étude exploratoire, repose sur un volume de corpus important, mais dont le format reste spécifique. Les tweets avec un nombre restreint de caractères imposés, sont particulièrement propices à la CHD (qui se base sur des segments de textes), mais ne laissent que peu de place pour la parole de leur auteur. L'exploration manuelle de nombreux tweets sur chaque classe, nous a permis d'observer une forte tendance à reprendre les titres ou mots clés des articles cités. Cette pratique amoindrit la place accordée à l'argumentaire de l'auteur du tweet. Ces caractéristiques étant constitutives de notre terrain, les résultats obtenus sont à l'image de ces pratiques. Afin d'affiner notre travail, et d'explorer plus précisément la parole propre des auteurs de tweets citant OpenEdition, une étude supplémentaire devra être portée, sur les éléments indépendants des contenus cités par exemple.

Au-delà de ces limites, cette étude si elle ne peut être exhaustive, nous permet de relever certaines tendances. Un premier découpage des contenus textuel est effectué sur un critère linguistique. Avec une majorité de tweets francophone, les contenus hispaniques et germanophones se sont naturellement agrégés en classes aux lexiques homogènes. L'étude des relations entre discours et clusters de retweets sur les contenus germanophones, montre une forte corrélation entre la langue du compte Twitter et celui des contenus de tweets. Si le discours des tweets anglophones est également agrégé en deux classes, il affiche une hétérogénéité quant aux clusters des comptes ayant émis ces tweets. Ce trait d'universalité est le résultat de l'usage de l'anglais comme langue de la science. L'analyse des tweets francophones montre qu'ils respectent un cloisonnement disciplinaire où les comptes Twitter ont tendance à retweeter les contenus de leurs disciplines.

Dans chacune de ces sections nous observons la diffusion des nouvelles publications et des actualités de ces communautés scientifiques (appel à communication, programme de colloque...). L'exploration des tweets nous a également permis de mettre en relief un usage des contenus d'OpenEdition dans une visée formative. Ainsi, si la thématique éducative constitue une classe spécifique, des pratiques de recommandations de lecture ressortent sur les autres classes de discours avec un objectif pédagogique, que ce soit pour la préparation de cours ou la préparation aux concours de recrutement d'enseignants. Enfin, les tweets

contenant un lien vers OpenEdition au-delà des contenus disciplinaires hébergés, font apparaître une forte proportion de discours concernant la science ouverte. Elle s'exprime ici sous plusieurs domaines, à la fois terrain d'étude, courant de pensée, enjeux sociétal mais aussi domaine économique.

**Ce travail a bénéficié d'une aide du gouvernement français au titre du Programme Investissement d'Avenir, Initiative d'Excellence d'Aix-Marseille Université – A\*MIDEX**

## NOTES

[1] <http://mashe.hawksey.info/twitter-archive-tagsv5/>

[2] Revue.org jusqu'en 2017

[3] Les chi2 de liaison entre tweets provenant du cluster germanique et tweets appartenant aux classes sont de 83 220 pour la classe 15 et de 107 006 pour la classe 16

## REFERENCES BIBLIOGRAPHIQUES

BASTIAN Mathieu, HEYMAN, Sébastien, & JACOMY Mathieu (2009). Gephi: an open source software for exploring and manipulating networks. *Icwsn*, 2009, n°8, p.361-362

BLONDEL Vincent, GUILLAUME Jean-Loup, LAMBIOTTE Renaud, & LEFEBVRE, Etienne. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008, n°10, P10008

COSTAS Rodrigo, VAN HON Koen, FRANSSSEN Thomas. (2017). *Scholars on Twitter: who and how many are they?* arXiv preprint arXiv:1712.05667

HICKS Diana., WOUTERS Paul., WALTMAN Ludo., RIJKE Sarah, RAFOLS Ismael. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature Publishing Group*, 2015, vol. 520(7548), p.429-431

IBEKWE-SANJUAN Fidelia, BOWKER Geoffrey (2017). Implications of big data for knowledge organization. *Knowledge Organization*, 2017, vol. 44, n°3, p.187-198

LOUBERE Lucie. (2014). *Le traitement des TICE dans les discours politiques et dans la presse*. Présenté à 12èmes Journées internationales d'Analyse statistique des Données Textuelles

MAYAFFRE Damon. (2008). L'entrelacement lexical des textes. Cooccurrences et lexicométrie. *Texte et corpus*, 2008, n°3, Actes des Journées de la linguistique de Corpus 2007, p.91-102

O'NEIL Jill. (2016). NISO Recommended Practice: Outputs of the Alternative Assessment Metrics Project. *Collaborative Librarianship*, vol. 8, n°3

RATINAUD Pierre. (2009). *IRaMuTeQ: Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires*. Téléchargeable à l'adresse: <http://www.iramuteq.org>.

RATINAUD Pierre, MARCHAND Pascal. (2015). Des mondes lexicaux aux représentations sociales. Une première approche des thématiques dans les débats à l'Assemblée nationale (1998-2014). *Mots. Les langages du politique*, n°108, p.57-77

REINERT Max. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 1993, n°66, p.5-39

REINERT Max. (1983). Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, t. 8, n°2, p.187-198

ROBINSON-GARCIA Nicolás, SUGIMOTO Cassidy, MURRAY Dakota, YEGROS-YEGROS Alfred, LARIVIERE Vincent, COSTAS Rodrigo. (2019). The many faces of mobility: Using bibliometric data to measure the movement of scientists. *Journal of Informetrics*, 2019, vol. 13, n°1, p.50-63

UNESCO. (s. d.). *Libre accès à l'information scientifique*. Organisation des Nations Unies pour l'éducation, la science et la culture. Consulté 2 janvier 2019, à l'adresse <http://www.unesco.org/new/fr/communication-and-information/access-to-knowledge/open-access-to-scientific-information/>