L'ouverture des données de la recherche dans le cadre d'un projet pluridisciplinaire entre SIC et informatique : le cas des médias sociaux de santé

*Open Research Data in a Multidisciplinary Project Combining Information and
Communication Science and Computer Science. The Case of Social Media in
Healthcare*

Viviane Clavier et Céline Paganelli

Édition électronique

URL : <http://journals.openedition.org/edc/8759>

DOI : 10.4000/edc.8759

ISSN : 2101-0366

Éditeur

Université de Lille

Édition imprimée

Date de publication : 1 juin 2019

Pagination : 117-136

ISBN : 978-2-917562-21-5

ISSN : 1270-6841

Référence électronique

Viviane Clavier et Céline Paganelli, « L'ouverture des données de la recherche dans le cadre d'un projet pluridisciplinaire entre SIC et informatique : le cas des médias sociaux de santé », *Études de communication* [En ligne], 52 | 2019, mis en ligne le 01 janvier 2021, consulté le 21 janvier 2021. URL : <http://journals.openedition.org/edc/8759> ; DOI : <https://doi.org/10.4000/edc.8759>

L'ouverture des données
de la recherche dans le cadre
d'un projet pluridisciplinaire entre SIC
et informatique : le cas des médias
sociaux de santé

*Open Research Data
in a Multidisciplinary Project Combining
Information and Communication
Science and Computer Science.
The Case of Social Media in Healthcare*

Viviane Clavier

Université Grenoble-Alpes, Gresec EA 608, F-38130 Échirolles, France
viviane.clavier@u-grenoblealpes.fr

Céline Paganelli

Université Paul Valéry Montpellier 3, LERASS EA 827,
F-34000 Montpellier, France
celine.paganelli@univ-montp3.fr

Résumé / Abstract

La question de l'ouverture des données est posée dans les recherches portant sur les médias sociaux. Dans un premier temps, nous identifions la nature des données, des méthodes et des traitements mobilisés dans les articles de deux revues qualifiantes en SIC afin d'observer quelle place est accordée aux données, une étape préalable à leur éventuelle ouverture. Dans un second temps, nous prenons appui sur une expérience critique issue d'un programme de recherche sur la prévention du VIH consacré à la détection de conduites à risques sur les médias sociaux. Nous mettons en évidence des questionnements liés au cycle des données et aux données personnelles de santé.

Mots-clés : ouverture des données, données de la recherche, médias sociaux, santé, plateforme d'annotation.

We explore the issues surrounding the opening of social media research data. Firstly, in order to observe what place is given to data in social media research, we analyse articles published in two prominent French Information and Communication Science journals. This analysis allows us to identify and to describe the nature of the data, the methods used and the type of processing mentioned in articles dealing with social media research. Secondly, we discuss a critical experience stemming from a research program on HIV prevention dedicated to the detection of risky behavior on social media. In our discussion, we highlight questions related to the data cycle and to personal health data.

Keywords: open data, research data, social media, health, annotation platform.

Introduction

Les données de la recherche sur lesquelles nous travaillons sont issues de médias sociaux de santé, se présentent comme massives et dégradées, sont publiques tout en étant intimes.

En nous appuyant sur des exemples puisés lors de la réalisation d'un programme de recherche pluridisciplinaire sur les médias sociaux et la prévention du VIH (2016-2017)¹, nous proposons un retour d'expérience complémentaire d'études et enquêtes sur l'ouverture des données en SHS (Prost et Schöpfel, 2015 ; Serres et al., 2017). Ces enquêtes mesurent l'écart entre les discours d'injonction ou de recommandation avec les pratiques réelles et les représentations des chercheurs (Serres et al., 2017, 7). Contrairement aux archives ouvertes, les données constituent encore largement un « impensé » (Serres et al., 2017, 41), mais leur ouverture se présente pourtant comme l'un des volets de ce qu'il convient désormais d'appeler une « politique » de l'open access et non plus un « mouvement » en raison d'un « cadrage politique au plus haut niveau de l'État » et de la mise en place de mesures incitatives, voire coercitives, de la part des organismes financeurs de la recherche (Chartron, 2016).

L'ouverture des données n'est certes pas une question nouvelle, cependant les objectifs formulés par les instances scientifiques ont évolué depuis la *Déclaration de Berlin* en 2004. D'objectifs scientifiques préconisant la validation d'hypothèses par des données probantes à la réutilisation des données, les données endossent depuis 2016 les vertus de transparence financière et de démocratie². Enfin, le mouvement de la Science Ouverte, repris par le CNRS, souligne des vertus morales d'intégrité scientifique³. Pour les SHS en particulier, qui jouissent d'une « valorisation omniprésente » en raison d'un large spectre d'interventions dans les champs sociaux, et paradoxalement d'une piètre visibilité⁴, l'ouverture des données représente, pour les instances de politique scientifique, un enjeu supplémentaire. Elle pourrait en effet constituer une opportunité de justifier l'« utilité sociale » (Gaudin, 2008) de ces

- 1 *Projet #AIDS Analyse Information Dangers Sexualité : détecter les conduites à risque d'exposition au VIH dans les médias sociaux*, Unités de recherche partenaires : LERASS-CERIC (Université Paul-Valéry, Montpellier), le GRESEC, (Université Grenoble-Alpes) et le LIRMM, (Université Montpellier).
- 2 L'article 30 de la loi du 7 octobre 2016 *Pour une République numérique* stipule que la réutilisation des données de la recherche doit être libre à partir du moment où la recherche a été financée au moins pour moitié par « [...] des dotations de l'État, des collectivités territoriales, des établissements publics, des subventions d'agences de financement nationales ou par des fonds de l'Union européenne [...] ».
- 3 CNRS, COMETS, *Guide pour une pratique de la recherche intègre et responsable*, CNRS, septembre 2016.
- 4 « *Paradoxe des SHS : une 'valorisation' à la fois omniprésente et manquant de visibilité* », [En ligne] disponible sur le site du CNRS à l'adresse : <http://www.cnrs.fr/inshs/recherche/actions-propres-institut/rtp/RTP%20Valorisation/pre-sentation.htm> (page consultée le 13 janvier 2019).

disciplines, qui ne peuvent se prévaloir de produire des brevets ou des licences d'exploitations quantifiables et objectivables sous la forme d'indicateurs. Or, cette ouverture est-elle toujours possible, voire souhaitable, et, s'agissant des médias sociaux, quelle est la plus-value d'ouvrir des données qui sont déjà en libre accès sur internet ?

Dans la suite, nous présentons les enjeux liés à l'analyse des données issues de médias sociaux, puis la méthode mise en œuvre dans le cadre du projet pluridisciplinaire sur le VIH dans les médias sociaux. Enfin, en nous appuyant sur l'expérience acquise au cours de ce projet, nous conduirons une réflexion autour de quatre questionnements soulevés par le recueil, le traitement, la mise à disposition et la diffusion des données issues de médias sociaux de santé.

1.

Enjeux des données issues des médias sociaux

1.1. La place des médias sociaux dans deux revues qualifiantes en SIC

Parmi les revues qualifiantes en Sciences de l'information et de la communication (SIC)⁵, nous avons choisi deux titres qui accordent une position centrale aux dispositifs sociotechniques afin d'observer quelle place occupent les recherches sur les médias sociaux. Il s'agit de la revue *Réseaux* créée en 1982, qui se présente comme la première revue française en information et communication et qui couvre un ensemble de disciplines⁶ et la revue *tic&société*, qui est fortement ouverte à la francophonie et analyse les rapports entre les technologies de l'information et de la communication (TIC) et la société⁷. Les deux revues révèlent une augmentation constante d'articles de recherche consacrés aux médias sociaux depuis une quinzaine d'années. C'est d'ailleurs à partir de 2008 que le nombre d'articles publiés sur ce sujet s'est accéléré (Graphique 1), le nombre d'articles entre 2002 et 2018 se montant à 95 (74 pour *Réseaux* et 21 pour *tic&société*) sur un total de 692 articles (579 pour *Réseaux* et 113 pour *tic&société*), soit un pourcentage de 13,7% pour la période considérée.

5 Disponible sur : <https://www.conseil-national-des-universites.fr/cnu/#/entite/entiteName/CNU/idChild/34> (page consultée le 13 janvier 2019).

6 Les disciplines représentées sont la sociologie, l'information-communication, les sciences politiques, la gestion, l'économie et l'histoire <https://www.cairn.info/revue-reseaux.htm?contenu=apropos> (page consultée le 13 janvier 2019).

7 <https://journals.openedition.org/ticetsociete/564> (page consultée le 13 janvier 2019).

Graphique 1 : Pourcentage d'articles de recherche qui traitent de médias sociaux dans les revues Réseaux et Tic&Société entre 2002 et 2018⁸

Sur les 95 articles recensés, les recherches portent majoritairement sur plusieurs médias sociaux (39 articles sur 95) et non sur un média en particulier à l'exception de Facebook et de Twitter, qui ont récemment fait l'objet de 8 et 6 publications. Les forums de discussion figurent parmi les dispositifs les plus cités – 15 articles sur 95 – devant les blogs (6 sur 95). Il faut noter qu'ils sont également mentionnés dans les études qui s'intéressent à des sites web, lesquels intègrent des forums. Les domaines ou thématiques sur lesquels portent les recherches sont variés, cependant la politique reste la plus présente, avec 16 articles sur 95⁹. Suivent ensuite 10 publications qui traitent des pratiques et des usages des médias socio-numériques dans le cadre professionnel, amateur ou dans la vie quotidienne. La santé – tout comme les sciences, la religion, les jeux vidéo en ligne ou les séries télévisées – est abordée dans 3 publications seulement.

Ces résultats s'appuient sur un corpus de revues limité. Toutefois, ils illustrent le positionnement scientifique adopté par les chercheurs en SIC qui étudient les médias en « s'attachant notamment aux phénomènes de médiation, de conception, de production et de réception, aux représentations, à

8 Les mots-clés utilisés pour interroger les bases d'articles tiennent compte de l'évolution des dénominations (réseaux sociaux, réseaux socio-numériques, forums, blogs, etc.), ainsi que des noms de réseaux sociaux comme Facebook, Twitter ou encore des sites web comportant des fonctionnalités du « web social », plus récemment les termes de plate-forme et de dispositifs se sont imposés et ont supplanté l'usage de l'acronyme Tics).

9 Selon Anaïs Théviot, le succès outre-atlantique de *mybarackobama.com* en 2012 a favorisé l'usage du web par les partis politiques, et depuis, plusieurs programmes de recherche ont vu le jour autour de cette thématique (Théviot, 2013).

l'appropriation des dispositifs sociotechniques et à l'innovation »¹⁰. Les médias sociaux occupent naturellement une place importante dans les études de la discipline, d'autant plus que l'accès facilité à ces dispositifs sur internet permet de recueillir des corpus. Les traces de comportements numériques représentent ainsi une « matière première » particulièrement profitable et de « nouvelles entités rendues accessibles par le numérique » (Boulier, 2014).

Dans les deux revues sélectionnées, les publications sur les médias sociaux apparaissent plus généralistes que spécialisées dans un domaine, à l'exception de la politique. Hormis les recherches sémio-discursives ou sémiotico-rhétoriques qui sont assez peu représentées, les médias sociaux sont interrogés selon les points de vue suivants.

- L'approche la plus répandue s'intéresse aux usages et aux pratiques info-communicationnelles. Elle s'attache à diverses « communautés », groupes, classes d'âge ou professions qui peuvent partager les mêmes loisirs, poursuivre les mêmes objectifs. Certaines études se limitent à un seul dispositif interactif (un Wiki, Twitter, Facebook, Flickr), d'autres considèrent un éventail plus large sans se limiter aux « nouveaux médias ».
- Une approche socio-économique des médias sociaux qui a pour ambition de définir plus rigoureusement la diversité des dispositifs, d'observer leur rôle, de comparer leurs usages prescrits et réels, d'étudier plus finement leurs spécificités ou leurs fonctionnalités en les replaçant dans un contexte macro-économique et socio-culturel plus large, d'identifier les acteurs présents sur les réseaux sociaux, et de comprendre leurs stratégies.
- Une approche centrée sur les médias sociaux considérés comme instruments d'une stratégie de communication (communication publique, communication interne ou externe, communication de prévention ou de risque) ou de marketing. L'objectif est d'observer les mutations de métiers ou d'activités spécifiques et d'identifier le rôle des dispositifs dans les organisations.
- Une approche orientée sur les contenus diffusés par les médias sociaux, qui conduit à caractériser l'information, à en observer les formes de production et d'usage ainsi qu'à étudier les modalités de diffusion des connaissances dans des domaines comme le sport, la santé, les sciences, etc. et à s'attacher à la circulation d'idéologies ou de discours.
- Enfin, une approche dite « critique » des médias sociaux qui, le plus souvent sous la forme d'essais (donc sans données) consiste à revenir sur des notions ou des concepts liés au développement des médias

10 Issu de la présentation de la 71^e section par le CNU : <https://www.conseil-national-des-universites.fr/cnu/#/entite/entiteName/CNU/idChild/34/id-Node/4471-4497> (page consultée le 13 janvier 2019).

sociaux (le concept de réseau social, de communauté, d'espace public habermassien, de spatialisation et de territoire).

1.2. En quoi les médias sociaux sont-ils des données de la recherche pour les SHS ?

Si l'exploitation des données issues des médias sociaux est emblématique des *Big Data*, les recherches en SIC abordent ces données d'une manière spécifique. Plus précisément, le caractère volumétrique des corpus n'est pas un attrait, à l'exception de certains domaines comme le marketing numérique ou la politique, qui s'intéressent aux comportements sociaux, à la prise de décision stratégique ou politique. Dans ces cas-là, les « mesures passives dites '*Big Data*' »¹¹ entrent dans l'arsenal des méthodes utilisées par les instituts de sondage à titre complémentaire de méthodes qualitatives ou quantitatives.

A *contrario*, les études en SIC s'appuient souvent sur des corpus peu volumineux, l'injonction d'usage des méthodes numériques liées aux *Big Data* étant source de critiques majeures en SHS (Boyd et Crawford, 2011 ; Rouvroy et Berns, 2013 ; Ganascia, 2015 ; Bullich et Clavier, 2018). Dans le corpus de publications, nous avons effectivement trouvé très peu de grands jeux de données, hormis lorsque les recherches ont été réalisées au sein de projets nationaux (de type ANR) et ont impliqué des laboratoires en informatique, comme c'est le cas notamment d'une publication sur Twitter où 5 883 659 tweets ont été analysés (Rieder, 2012). Pour la grande majorité des articles publiés, malgré leur taille modeste, les données mobilisées n'en sont pas moins « significatives » à défaut d'être « représentatives »¹². La question de la représentativité qui garantit l'objectivité des données est déplacée du côté de la subjectivité que certains chercheurs sont amenés à justifier. Alors que l'exhaustivité est décrite comme un leurre, la variété des données est considérée comme une alternative à la représentativité pour les SHS (Boullier, 2014).

11 Pour Laurence Stoclet, directrice générale déléguée d'Ipsos, « Les réseaux sociaux constituent une nouvelle source de données très intéressante mais dont la représentativité n'est pas toujours assurée. Les enquêtes resteront donc une source de données importante, mais nous les complétons par des mesures passives dites '*big data*' (écoute du web par exemple, ou applications qui suivent les habitudes de surf sur les mobiles, tablettes ou PC) mais aussi par des techniques qualitatives d'observation ou de mesure des émotions (*facial coding, biometrics...*) ». Disponible à l'adresse : www.lenouveleconomiste.fr/lesdossiers/instituts-detudes-et-de-sondages-big-data-highlighting-you-tou-31828/ (page consultée le 13 janvier 2019).

12 Comme l'indiquent Tomasso Venturini *et al.* (2014, 11), la représentativité n'est pas directement liée à la taille : « Il est illusoire de penser que la taille des données circulant sur les flux des API (*Application protocol interface*) de Google ou de Twitter serait suffisante pour en faire de fidèles miroirs de la vie collective. La théorie statistique enseigne en effet que la représentativité d'un échantillon ne dépend qu'indirectement de sa taille : la qualité d'un échantillon dérive plutôt de sa ressemblance à la population de référence, de sa capacité d'en inclure la même variabilité ».

Les deux enquêtes conduites dans les universités de Lille (Prost et Schöpfel, 2015) et de Rennes (Serres *et al.*, 2017) illustrent la diversité des données collectées par les chercheurs en SHS. Parmi les sources les plus utilisées, figurent les textes (72 % à Rennes et 64 % à Lille), les enquêtes et les entretiens (53 % à Rennes et 47 % à Lille) puis les observations (43 % à Rennes et 41 % à Lille)¹³. Nous avons observé des proportions similaires dans les revues *Réseaux* et *tic&société* pour les entretiens et les observations. En revanche, la notion de « textes », et celle de « corpus » est paradoxalement très minoritaire, hormis dans les approches inspirées du *Big Data*, pour lesquelles on peut supposer que les outils numériques ont permis d'aspirer les sites et de les stocker sous forme de corpus. Seules les vidéos semblent plus fréquemment téléchargées (« Le corpus que j'ai pu constituer [...] contient une trentaine d'extraits vidéo »). Ailleurs, les auteurs parlent de sites web, de forums, de blogs dont le nom est parfois énoncé, mais pas l'URL (« plusieurs fils de discussion, tout thème confondu, ont été analysés » ; « nous avons complété ces entretiens par une analyse de contenu de blogs de cinéma faisant l'objet de ces sollicitations »).

Ainsi, deux remarques s'imposent au sujet des données de la recherche telles qu'observées dans les deux revues que nous avons étudiées. D'une part, la culture de la donnée ne semble pas une priorité dans les publications étudiées qui ne mettent pas systématiquement en évidence les caractéristiques quantifiées et documentées des données recueillies et parfois même, de leur traitement (à l'exception de méthodes originales où le chercheur va alors développer les principes théoriques et méthodologiques). Quand elle est indiquée, la mention du matériau de la recherche est souvent peu visible, située à la fin de l'introduction ou très souvent, en note de bas de page. D'autre part, le fait que les données semblent rarement archivées – à l'exception des données d'usages ou des traces de connexion qui sont systématiquement extraites et enregistrées dans des bases de données –, contrevient à l'un des quatre principes¹⁴ d'une approche des données comme « matériel », à savoir « l'enregistrement comme préalable de la reconnaissance » (Prost et Schöpfel, 2015, 7). Or, le caractère éphémère de l'internet laisse malheureusement présager qu'en l'absence d'archivage, la plupart des sources sur lesquelles s'appuient les recherches aura probablement disparu dans les années qui suivent la publication.

1.3. Comment les médias sociaux sont-ils traités dans une approche de données massives ?

Une abondante littérature fait état de recherches qui confrontent les méthodes issues des sciences de la donnée avec les sciences humaines et

¹³ (Serres *et al.*, 2017, 26).

¹⁴ Ces quatre principes sont : l'enregistrement préalable, la nature factuelle des données, le lien avec la communauté scientifique et la finalité des données en vue de la validation des hypothèses (voir Schöpfel *et al.*, 2017, 3-4).

sociales (Anderson, 2008 ; Boulier, 2014 ; Bourdeloie 2014 ; Longhi, 2017). Les questionnements portent sur l'articulation entre les méthodes quantitatives, qualitatives et automatisées (Venturini *et al.*, 2014) ; ils s'attachent à la redistribution et la requalification des méthodes numériques et à leur application possible en SIC (Plantin et Monnoyer-Smith, 2013). Les discussions portent également sur l'apport de ces méthodes pour analyser le web politique (Boyardjian *et al.*, 2017, Théviot, 2016).

Concernant les traitements appliqués aux médias sociaux, la santé se présente comme un terrain d'étude particulièrement propice à observer. En effet, les médias sociaux de santé sont l'objet d'une attention privilégiée des instituts de sondage¹⁵ et des institutions de santé publique qui se focalisent sur les comportements à risques comme les risques liés au suicide (Luxton *et al.*, 2012), ou la prévention du suicide (Maigrot *et al.*, 2017). Les résultats de ces traitements numériques sont utilisés pour orienter les politiques publiques. Les méthodes s'attachent à la recherche de thématiques via des mots-clés (recherche de symptômes, de conduites à risques, de drogues, etc.) et recourent à des analyses de données textuelles ou des méthodes de fouille de textes.

Sur le VIH, des travaux ont été menés, notamment aux États-Unis, qui s'appuient sur une analyse des médias sociaux pour suivre l'épidémie du virus (Thangarajan *et al.*, 2015) ou repérer des indicateurs du risque liés au VIH (Weibel *et al.*, 2017). Dans cette dernière étude, les auteurs soulèvent les questions éthiques que fait émerger l'analyse de données sensibles issues des médias sociaux. Si l'interrogation n'est pas nouvelle et était déjà vive au début d'internet (Jones, 1994), elle prend une toute autre dimension dans un contexte où les données accessibles sont de plus en plus volumineuses et où les capacités technologiques pour l'analyse sont démultipliées (Weibel *et al.*, 2017).

Bien que les attentes sociales en matière de traitement de données massives soient pressantes, tous les outils à disposition n'ont pas été conçus dans le but de « servir » les recherches en sciences humaines et sociales. À l'exception de quelques domaines bien identifiés qui font usage depuis des décennies de méthodes assistées par ordinateur, telle la littérature ou la politique, d'autres méthodes, comme la classification automatique supervisée ou non supervisée, sont loin d'avoir des applications grand public. Ainsi, les recherches qui sont conduites par les chercheurs en informatique visent le développement et l'amélioration de méthodes davantage que celles d'outils, alors que les chercheurs en SIC qui travaillent sur les données s'intéressent aux méthodes et outils uniquement comme des instruments pour comprendre et analyser des phénomènes ou des situations info-communicationnelles. C'est pourquoi, la

15 Par exemple *Les Baromètres santé, un observatoire des comportements des Français pour orienter les politiques de santé publique*. Accessible sur <http://inpes.santepubliquefrance.fr/Barometres/index.asp> (page consultée le 13 janvier 2019).

collaboration scientifique pluridisciplinaire sur des projets mettant en jeu des données massives n'amène pas toujours les résultats escomptés et suscite des déceptions de part et d'autre, mettant également à jour des obstacles ou difficultés que le partenariat peine parfois à contourner. C'est aussi le cas lorsqu'il s'agit de valoriser les résultats de la recherche. En effet, les normes de publication scientifique sont différentes dans les deux disciplines : par exemple, publier à plusieurs ne pose aucun problème en informatique alors que cette pratique est peu reconnue en SIC.

2.

Un exemple : une méthodologie croisée pour repérer le risque et l'émotion dans un corpus sur le VIH

Nous avons participé, en 2017, à un projet de recherche intitulé *#AIDS Analyse Information Dangers Sexualité : détecter les conduites à risques via/ dans les médias sociaux* financé par l'ANRS¹⁶, réunissant des chercheurs en SIC et en informatique¹⁷, ainsi que des acteurs de la santé et de la prévention : le Corevih de Montpellier¹⁸ et Sida-Info-Service qui a mis à disposition les archives du forum de discussion de l'association. Le projet visait l'analyse des contenus, en langue française, échangés sur le forum de l'association et d'autres médias sociaux consacrés au VIH pour identifier les discours relatifs aux questions de sexualité, de consommation de produits illicites, et devait faire émerger les comportements à risque ainsi que les facteurs liés à la *prise de risque* (Clavier et Paganelli, 2017). L'analyse, combinant une démarche qualitative et quantitative, a conduit au développement de méthodes et d'outils permettant une semi-automatisation du traitement des données recueillies et la visualisation temporelle des résultats (Mercadier *et al.*, 2018). Une plate-forme d'annotation a été développée et présente actuellement plus de 4 000 messages annotés manuellement.

L'objectif était double. D'un point de vue opérationnel, il s'agissait de repérer, grâce à une approche méthodologique qualitative combinée à des outils de classification automatique, des comportements à risque ainsi que des facteurs conduisant à cette prise de risque. Le projet permettait également de s'interroger sur la manière dont les sciences humaines et sociales, et notamment les SIC d'une part, et l'informatique d'autre part, dialoguent pour

16 Agence Nationale de Recherche sur le SIDA. Nous remercions l'ANRS qui a financé ce projet dans le cadre des contrats d'initiation.

17 Jérôme Azé et Sandra Bringay, professeurs et chercheurs au Lirmm, Frédéric Assié, étudiant.

18 Le professeur Jacques Reynes et Magali Reynes, coordinatrice au sein du Corevih.

combiner leurs approches méthodologiques voire co-produire de nouvelles méthodes d'analyse et de traitement (Darbellay *et al.*, 2018).

2.1. La collecte des données

La première tâche, menée par les informaticiens, en lien avec Sida-Info-Service, a consisté à extraire les données de leur forum, à les structurer et à les prétraiter. À côté du forum de SIS, ont été recueillis des messages du forum « Sida et prévention » de *Doctissimo* ainsi que des tweets. Les données ont été structurées au sein d'une base de données. 228 235 messages ont été collectés, issus de 8 565 utilisateurs.

La constitution d'un corpus répond à un objectif de recherche et s'effectue dans le cadre de contraintes techniques, éthiques et légales (Rieder, 2012). Pour répondre à notre objectif qui consistait à faire émerger les comportements à risque ainsi que les facteurs liés à la *prise de risque*, nous avons besoin d'un corpus présentant plusieurs caractéristiques. Il se devait d'être volumineux pour permettre aux informaticiens d'entraîner leurs outils de classification automatique, récent pour identifier des facteurs émergents liés à la prise de risque, et varié de manière à observer d'éventuelles différences ou régularités entre plusieurs médias sociaux. Mais collecter un tel corpus se heurte à quelques obstacles. Nous en relèverons trois.

En premier lieu, le recueil de tweets n'a pas apporté les résultats attendus. Parmi les possibilités offertes par les API de Twitter (Rieder, 2012) et même si des restrictions peuvent rendre la collecte difficile (Courtin *et al.*, 2015), il est possible de recueillir des tweets selon leur auteur ou bien en filtrant par mots-clés. La collecte des messages produits par un ensemble d'utilisateurs prédéterminé n'a pas été choisie car elle n'était pas pertinente ici. Elle supposait que nous sachions de quels individus, organisations, communautés émanaient les messages susceptibles de nous intéresser. Les informaticiens ont donc programmé une collecte de tweets comportant des termes en lien avec le virus et la prise de risque (VIH, sida, sexe, drogues). Les outils ont permis de recueillir rapidement des données nombreuses mais la pertinence de la collecte est discutable. Le filtrage par mots-clés a ses limites : il est difficile de dire si les messages collectés sont représentatifs de l'ensemble des tweets sur le sujet ou si d'autres termes n'auraient pas été utilisés dans les échanges sur le sujet. Les problèmes récurrents des études sur la langue se posent : synonymie, polysémie. Par ailleurs, un grand nombre de messages recueillis ont à voir avec la prévention et ne reflètent pas des comportements d'internautes. En mettant en relation les auteurs et les contenus, nous aurions pu isoler les messages émanant des pouvoirs publics ou des associations de prévention par exemple.

Ensuite, se posent des questions éthiques liées à la diffusion de la plateforme. D'une part, l'exploitation des données issues de la recherche apparaît comme sensible à toutes les étapes de la récupération, de leur exploitation et de leur diffusion. D'autre part, les données issues des médias sociaux sont spé-

cifiques lorsqu'elles abordent des contenus liés à l'intimité. Ainsi, les données collectées, sans pour autant être personnelles – elles ont été anonymisées –, n'en sont pas moins intimes. Cette spécificité incite à une certaine prudence dans la confiance que l'on peut accorder au matériau de la recherche. Des auteurs considèrent en effet qu'internet, sur les sujets de la sexualité, aurait « une fonction de défoulement » et serait révélateur de valeurs et d'interactions sociales (Dannecker, 2002). Certains indiquent même qu'il y aurait moins de risques de distorsions concernant l'identification des comportements à risques lorsqu'elles sont conduites dans l'anonymat d'internet qu'en face à face (Binik *et al.*, 1999). D'autres au contraire doutent de la fiabilité des discours produits sur les réseaux socio-numériques : c'est le cas des *datascientists* qui accordent davantage de crédit aux conduites réelles et aux traces d'activité des internautes qu'à leurs allégations pour fonder leurs méthodes d'apprentissage (Cardon, 2015, 34).

Le recueil d'un corpus de données sur internet se heurte enfin à des questions juridiques. Sur ce point, le cadre est, pour l'instant, encore flou. Malgré la loi *Informatique et libertés* de 1978 et l'entrée en application en mai 2018 du RGPD (*Règlement général sur la protection des données personnelles*) au niveau européen, aucun service juridique de nos différentes universités ne nous a donné les mêmes informations sur la marche à suivre. Dans la mesure où les données sont accessibles librement sur internet, avons-nous le droit de les récupérer et de faire porter des traitements (qu'ils soient manuels ou automatisés) sur le corpus recueilli ? La seule contrainte est-elle d'anonymiser ou bien des autorisations sont-elles à demander ? Et à qui ? Aux internautes ? Mais dans ce cas, le nombre de personnes concernées et la difficulté pour les contacter rend la tâche impossible, à l'organisme (qu'il soit public ou privé) qui gère le site ? Finalement, sur ce point, un grand nombre d'interrogations subsistent auxquelles la gestion du projet n'a pas permis de répondre, mais qui sont aujourd'hui posées ouvertement au sein de la communauté scientifique (Ginouves et Gras, 2018).

2.2. Les traitements

La deuxième tâche, menée par les chercheurs en SIC, consistait à catégoriser les messages. Au sein du corpus, nous avons annoté 4 413 messages issus de 64 fils. Par ailleurs, nous avons circonscrit un sous-corpus composé d'un fil de discussion de 261 messages, sur lequel une analyse fine relative au risque et à l'émotion a été menée. L'annotation manuelle reposait sur trois niveaux d'analyse. Au premier niveau, l'analyse de contenu visait à caractériser la dynamique des interactions et à typer les informations véhiculées à partir d'une grille testée précédemment sur des forums de santé (Paganelli et Clavier, 2011). La catégorisation des productions textuelles a pris en compte deux éléments : l'intentionnalité de leurs auteurs et la nature de l'information véhiculée (Clavier *et al.*, 2010). Au deuxième niveau, une analyse de contenu et une analyse des discours rapportés ont mis en évidence des éléments de

signification sur le compte d'une axiologie de la confiance et de la défiance avec le corps médical et d'une axiologie de la proximité ou de la dépersonnalisation du soin. Et enfin, le troisième niveau était consacré à la manière dont le risque d'exposition au VIH était vécu par les internautes. Les récits, souvent empreints d'émotions, peuvent renvoyer à une situation factuelle passée, présente ou encore à venir. En nous appuyant sur le modèle de Raphaël Micheli (2013), nous avons annoté 261 messages selon 3 types d'émotions : l'émotion dite, l'émotion montrée et l'émotion étayée. Ces annotations ont permis aux collègues informaticiens de développer, dans une tâche suivante, une méthode semi-automatique en s'appuyant sur les caractéristiques des messages annotés manuellement (Mercadier *et al.*, 2018).

Le traitement des données, dans le cadre d'un projet associant plusieurs disciplines suscite également des interrogations.

En premier lieu, le traitement de données récupérées sur internet implique de se poser la question de la protection des données. L'accès à ces dernières nécessite une anonymisation. En effet, une distinction est faite entre anonymisation et pseudonymisation. Les données ne sont pas anonymes lorsqu'il demeure possible d'individualiser une personne en recoupant les données collectées avec d'autres données ou lorsque le responsable de traitement peut remonter le processus de chiffrage (selon la décision du conseil d'État de février 2017).

Ainsi, lors de la tâche 1, une étape d'anonymisation a été appliquée afin de rendre impossible la « ré-identification » des personnes ou des entités. Les noms et pseudonymes ont été remplacés par des étiquettes, les dates modifiées. Cela pose au moins deux problèmes. Le premier réside dans la décontextualisation des propos recueillis. Puisqu'il n'est plus possible par exemple d'accéder aux informations liées au pseudonyme, nous étudions les messages de manière décontextualisée. Toutes les informations que l'internaute souhaite partager via son pseudo sont perdues (âge, sexe, orientation sexuelle par exemple). En outre, il n'est pas possible de relier entre eux les messages postés précédemment par le même rédacteur. Par ailleurs, l'anonymisation ne concerne pas seulement les émetteurs des messages. Ainsi, tous les noms propres ayant été remplacés par des étiquettes, nous perdons des informations en lien avec des tests de dépistage, des noms de centres de soins, de laboratoires pharmaceutiques qui pourraient, s'ils reviennent de manière fréquente, être utiles lors d'une analyse en lien avec la prise de risque. Enfin, comme les données ont été rendues anonymes et que le contexte d'usage de l'information relève de la sphère intime, il est difficile de recourir à des entretiens de personnes inscrites sur ces forums, alors qu'une des forces des approches méthodologiques en SIC réside souvent dans le croisement de données issues d'une analyse de corpus et de discours recueillis lors d'entretiens.

Un second questionnement soulevé par le projet est lié au changement d'échelle en matière de traitement. Travailler sur des corpus volumineux né-

cessite la mise en place d'outils et de procédures qui étaient, le plus souvent, étrangers aux chercheurs en SIC¹⁹. S'ils facilitent l'annotation et l'exploitation des corpus, ces outils impliquent une rationalisation des *manières de faire* et parfois une simplification des traitements, en raison de la diversité des objectifs attachés aux enjeux de l'annotation. Pour les informaticiens, le corpus d'entraînement sera toujours trop petit, et les annotations toujours trop fines. Pour les chercheurs en SIC, l'annotation n'est pas un objectif en soi, le corpus annoté n'ayant pas de valeur intrinsèque et les méthodes utilisées étant liées à un objectif scientifique qui varie suivant la recherche à conduire. Ainsi, le changement d'échelle n'est pas compatible avec l'annotation manuelle et les méthodes qualitatives sont difficilement applicables dans un contexte de traitement de données massives.

2.3. L'ouverture des données

Dans le cadre du projet, les informaticiens ont développé une plate-forme, accessible sur internet, qui rassemble les données recueillies et propose des fonctionnalités d'annotation. Dans l'idéal, cet outil devrait être à disposition d'autres chercheurs comme des acteurs de la prévention du soin, de manière à donner accès aux données annotées. Dans les faits, le choix des données, des modalités de mise à disposition n'est pas simple et nécessite une connaissance du cadre juridique et réglementaire des projets d'*open data* (Castets-Renard et Gandon, 2016). Par exemple si les données n'ont pas été totalement créées par le chercheur mais proviennent d'un tiers, comme c'est le cas dans notre projet, un cadre contractuel est nécessaire. Qu'en est-il lorsque les données proviennent d'internet ? Le choix de la contextualisation des données est nécessaire également. Au-delà de la plate-forme d'annotation, outil de travail du chercheur, quelles métadonnées rajouter aux données notamment en vue des acteurs de la prévention ? Enfin quel serveur d'hébergement privilégier lorsque le projet est multi-partenaires ? Et quel cadre juridique choisir concernant la réutilisation des données ? Les partenaires d'un projet se trouvent souvent bien démunis face à ces questions qui nécessiteraient un traitement au niveau des institutions. Certains organismes de recherche proposent, par exemple, des chartes ou guides de l'*open data*²⁰.

Discussion et conclusion

En nous appuyant sur les articles de deux revues qualifiantes, nous avons analysé les données et les méthodes mentionnées dans les publications des

19 Sauf la petite communauté de collègues en SIC qui travaillent avec des chercheurs en TAL.

20 C'est le cas de l'Inra par exemple : <https://www6.inra.fr/datapartage/> (page consultée le 13 janvier 2019).

chercheurs. Cette entrée par les discours devrait être complétée par des entretiens afin d'apprécier, dans les cas où les données et méthodes sont peu ou pas développées, s'il s'agit d'une volonté délibérée d'effacer toute trace d'empirisme, ou tout simplement une absence. S'agissant de notre expérience, celle-ci est loin d'être un cas isolé, les SIC étant fréquemment sollicitées dans des projets pluridisciplinaires. Elle a permis de conforter nos méthodes d'annotation manuelle, de valider notre grille d'analyse, et d'évaluer notre système de catégorisation en vue d'entraîner des données. Grâce à la plate-forme d'annotation développée par les informaticiens, l'annotation manuelle a pu être effectuée de manière plus efficace, plus rapide, et a permis un travail partagé. Concernant l'automatisation, selon nos collègues informaticiens, « Les résultats sont prometteurs et montrent l'efficacité des méthodes d'apprentissage profond pour caractériser les messages des forums de manière automatique. La méthode de visualisation mise en place permet d'explorer les résultats de ces méthodes et ainsi faciliter l'accès aux connaissances » (Mercadier *et al.*, 2018). Notre collaboration a permis un enrichissement de nos approches et méthodes mutuelles, dont l'enchaînement est cependant surtout resté séquentiel. Il nous faudrait à présent appliquer ces méthodes à d'autres corpus portant sur des thématiques proches (les addictions) ou d'autres dispositifs, faisant intervenir des populations plus jeunes (le *chat* de SIS).

Pour terminer, alors que les injonctions des pouvoirs publics et les attentes des agences d'évaluation de la recherche sont fortes en matière d'ouverture des données, les politiques et procédures mises en œuvre au sein des universités, sont encore peu formalisées, voire inexistantes, et rendent complexe le parcours des chercheurs engagés dans ce type de projet.

Bibliographie

Anderson C. (2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". In *Wired*. Disponible sur : <http://www.wired.com/2008/06/pb-theory/> (page consultée le 13 janvier 2019).

Binik Y. M., Mah K., Kiesler S. (1999). "Ethical issues in conducting sex research on the internet". In *The Journal of Sex Research*, vol. 36, n° 1, p. 82-90.

Boullier D. (2014). *Les sciences sociales face aux traces du big data ? Société, opinion et répliques*. FMSH-WP-2015-88. Disponible sur : <https://halshs.archives-ouvertes.fr/halshs-01141120/> (page consultée le 13 janvier 2019).

Bourdaloie H. (2014). « Ce que le numérique fait aux sciences humaines et sociales. Epistémologie, méthodes et outils en questions ». In *tic&société*, vol. 7, n° 2, p. 7-38.

Boyd D., Crawford K. (2011). "Six Provocations for Big Data". In *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, September. Disponible sur SSRN : <https://ssrn.com/abstract=1926431> (page consultée le 13 janvier 2019).

Boyadjian J., Olivesi A., Velcin J. (2017). « Le web politique au prisme de la science des données. Des croisements disciplinaires aux

renouvellements épistémologiques ». In *Réseaux*, vol. 204, n° 4, p. 9-31.

Bullich V., Clavier V. (2018). Dossier « Production des données, 'production de la société'. Les Big Data et algorithmes au regard des sciences de l'information et la communication ». In *Les Enjeux de l'information et la communication*, vol. 19, n° 2, p. 5-177. Disponible en ligne sur : <https://lesenjeux.univ-grenoble-alpes.fr/pageshtml/art2018.html#dossier> (page consultée le 13 janvier 2019).

Cardon D. (2015). *A quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Seuil et La République des Idées.

Castets-Renard C., Gandon N. (2016). « Open data des données de la recherche publique : entre réformes législatives et retour d'expérience sur un guide pratique à destination des chercheurs ». In *LEGICOM*, vol. 56, n° 1, p. 67-75.

Chartron G. (2016). « Stratégie, politique et reformulation de l'open access ». In *Revue française des sciences de l'information et de la communication*, vol. 8. Disponible sur : <http://journals.openedition.org/rfsic/1836> (page consultée le 13 janvier 2019).

Clavier V., Manes-Gallo M.-C., Mounier E., Paganelli C., Romeyer H., Staii A. (2010). Ch. 19 « Dynamiques interactionnelles et rapports à l'information dans les forums de discussion médicale ». In Mille-
rand F., Proulx S., Rueff J. (dir.), *Le web social : mutation de la*

communication, Québec, Presses Universitaires du Québec, p. 297-312.

Clavier V., Paganelli C. (2017). « Une approche méthodologique croisée du traitement des données de la recherche : le cas d'un corpus d'échanges issus de médias sociaux dans le domaine de la santé ». In *Colloque COSSI 2017, Méthodes et stratégies de gestion de l'information par les organisations : des "Big Data" aux "Thick Data"*, Université McGill, Montréal, Canada, 11-12 mai.

Courtin A., Foucault N., Juanals B., Licht C., Minel J.-L., Rampini M. (2015). « L'événement culturel #MuseumWeek : Projet ComNum Médiation culturelle et circulation des connaissances à l'ère numérique ». In *Rencontres Numériques : Médiation & numérique dans les équipements culturels*, Octobre, Créteil. Disponible sur : <https://hal.archives-ouvertes.fr/halshs-01218640> (page consultée le 13 janvier 2019).

Dannecker M. (2002). *Prévention du VIH et sexualité*. Disponible sur : <http://www.aidsnet.ch/modules.php?name=News-&file=print&sid=1554> (page consultée le 19 janvier 2019).

Darbellay F., Vinck D., Cocco C., Dessart G., Dandarova Z., Brandt P.-Y. (2018). « L'interdisciplinarité en partage : collaborer pour innover ». In *innovatiO*, vol. 5, *L'interdisciplinarité en action au sein des projets de recherche en innovation*. Disponible sur : <http://innovacs-innovatio.upmf-grenoble.fr/index.php?id=437> (page consultée le 13 janvier 2019).

Ganascia J.-G. (2015). « Les big data dans les humanités ». In *Critique*, vol. 819-820, n° 8, p. 627-636.

Gaudin J.-P. (2008). « Modalités et enjeux de l'évaluation des SHS : le tournant actuel ». In *Revue européenne des sciences sociales*, vol. XLVI-141. Disponible sur : <http://journals.openedition.org/ress/150> (page consultée le 13 janvier 2019).

Ginouves V., Gras I. (2018). *La diffusion numérique des données en SHS. Guide des bonnes pratiques éthiques et juridiques*, Aix-en-Provence, Presses universitaires de Provence.

Jones R. A. (1994). "The Ethics of Research in Cyberspace". In *Internet Research*, vol. 4, n° 3, p. 30-35.

Longhi J. (2017). « Humanités, numérique : des corpus au sens, du sens aux corpus ». In *Questions de communication*, vol. 31, n° 1, p. 7-17.

Luxton D. D., June J. D., Fairall J. M. (2012). "Social Media and Suicide: A Public Health Perspective". In *American Journal of Public Health*, vol. 102, n° 2, p. 195-200.

Maigrot C., Bringay S., Azé J. (2017). "Concept drift vs suicide: How one can help prevent the other?". In *International Journal of Computational Linguistics and Applications*, vol. 8, n° 1. Disponible sur : <https://hal.archives-ouvertes.fr/hal-01617870> (page consultée le 13 janvier 2019).

Mercadier Y., Azé J., Bringay S., Clavier V., Cuenca E., Paganelli C., Poncelet P., Sallaberry A. (2018).

« #AIDS Analyse Information Dangers Sexualité : caractériser les discours à propos du VIH dans les forums de santé ». In *IC 2018 : 29^e Journées francophones d'Ingénierie des Connaissances (Proceedings of the 29th French Knowledge Engineering Conference)*, Nancy, 4-4 juillet, p. 71-86. Disponible sur : <https://hal.archives-ouvertes.fr/hal-01839552> (page consultée le 13 janvier 2019).

Micheli R. (2013). « Esquisse d'une typologie des différents modes de sémiotisation verbale de l'émotion ». In *Semen*, vol. 35. Disponible sur : <http://semen.revues.org/9795> (page consultée le 13 janvier 2019).

Paganelli C., Clavier V. (2011). « Le forum de discussion : une ressource informationnelle hybride entre information grand public et information spécialisée ». In Ysari-Labrique E. (dir.), *Les forums de discussion : agoras du XXI^e siècle ? Théories, enjeux et pratiques discursives*, Paris, L'harmattan, p. 39-55.

Plantin J.-C., Monnoyer-Smith L. (2013). « Ouvrir la boîte à outils de la recherche numérique ». In *tic&société*, vol. 7, n° 2. Disponible sur : <http://journals.openedition.org/ticetsociete/1527> (page consultée le 26 février 2019).

Prost H., Schöpfel J. (2015). *Les données de la recherche en SHS. Une enquête à l'Université de Lille 3*. Rapport final. Villeneuve d'Ascq, Université de Lille 3. Disponible sur <http://hal.univ-lille3.fr/hal-01198379v1> (page consultée le 13 janvier 2019).

Rieder B. (2012). «The refraction chamber: Twitter as sphere and network». In *First Monday*, vol. 17, n° 11. Disponible sur : <http://firstmonday.org/ojs/index.php/fm/article/view/4199/3359> (page consultée le 13 janvier 2019).

Rouvroy A., Berns T. (2013). « Gouvernamentalité algorithmique et perspectives d'émancipation : Le disparate comme condition d'individuation par la relation ? ». In *Réseaux*, vol. 177, n° 1, p. 163-196.

Schöpfel J., Kergosien E., Prost H. (2017). « Pour commencer, pourriez-vous définir 'données de la recherche' ? Une tentative de réponse », In *Atelier VADOR : Valorisation et Analyse des Données de la Recherche*, INFORSID 2017, 31 mai, Toulouse.

Serres A., Malingre M.-L., Mignon M., Pierre C., Collet D. (2017), *Données de la recherche en SHS. Pratiques, représentations et attentes des chercheurs : une enquête à l'Université Rennes 2, Rapport de recherche*, Rennes, Université Rennes 2.

Thangarajan N., Green N., Gupta A., Little S., Weibel N. (2015). "Analyzing social media to characterize local HIV at-risk populations". In *WH'15: Proceedings of the Conference on Wireless Health*, Bethesda, Maryland- 14-16 octobre.

Théviot A. (2013). « Devenir 'ami' avec 4500 enquêtés. Les enjeux éthiques de l'analyse d'interfaces semi-privées ». In *tic&société*, vol. 7. Disponible sur <http://journals>.

openedition.org/ticetsociete/1608
(page consultée le 13 janvier 2019).

Théviot A. (2016). « Les data : nouveau trésor des partis politiques ? Croyances, constitutions et usages comparés des données numériques au Parti Socialiste et à l'Union pour un Mouvement Populaire ». In *Politiques de communication*, vol. 6, n° 1, p. 137-166.

Venturini T., Cardon D.,
Cointet J.-P. (2014). « Présentation ». In *Réseaux*, vol. 6, n° 188, p. 9-21.

Weibel N., Desai P., Saul L., Gupta A., Little S. (2017). "HIV Risk on Twitter: The Ethical Dimension of Social Media Evidence-based Prevention for Vulnerable Populations". In *HICSS'17: Proceedings of the 50th Hawaii International Conference on System Sciences*. Disponible sur http://aisel.aisnet.org/hicss-50/dsm/critical_and_ethical_studies/3
(page consultée le 13 janvier 2019).