HEALTHCARE
DATA INSTITUTE

INTERNATIONAL THINK TANK
DEDICATED TO BIG DATA IN HEALTHCARE

LES RÉSEAUX SOCIAUX ET LA SANTÉ :

UN ENJEU POUR LE SUIVI DES PATIENTS ET LA RECHERCHE SCIENTIFIQUE

SEPTEMBRE 2018

CO-AUTEURS (PAR ORDRE ALPHABÉTIQUE) :

- **Lina Autelitano,** Chef de projet digital, Direction Digitale, Pierre Fabre Médicament & Santé, membre du Healthcare Data Institute
- **Caroline Henry,** Avocat Associé, Pons & Carrère, membre du conseil d'administration du Healthcare Data Institute
- **Adel Mebarki,** Directeur général de Kap Code, membre du Healthcare Data Institute
- **Patrick Olivier,** Directeur général IVBAR France, Vice-Président du Healthcare Data Institute
- **Francisco Orchard,** data scientist Epiconcept, membre du Healthcare Data Institute

Coordination des travaux : Jean-Baptiste Fantun

LES RÉSEAUX SOCIAUX ET LA SANTÉ :

UN ENJEU POUR LE SUIVI DES PATIENTS ET LA RECHERCHE SCIENTIFIQUE

PRÉSENTATION

Convaincu que l'exploitation des données de santé ne peut se faire **que dans l'intérêt des patients, mais aussi avec les patients**, le Healthcare Data Institute a initié, au début de l'année 2017, une réflexion sur les patients et leurs données. Au mois d'avril 2017, une réunion de lancement était organisée avec des patients et représentants d'associations de patients. Les participants ont échangé sur l'accès des patients à leurs données de santé et le partage de ces données. Ces échanges ont montré que si le partage de données avec des équipes de recherche pouvait susciter des interrogations ou des appréhensions, le partage de ces mêmes données avec une communauté d'amis, de famille ou de patients sur les réseaux sociaux était en revanche une pratique fréquente et spontanée.

Partant de ce premier constat, le Healthcare Data Institute a décidé de structurer un groupe de travail pluridisciplinaire pour analyser les enjeux et perspectives d'usage des données générées sur les réseaux sociaux en santé[1]. Le groupe de travail a débuté ses travaux en septembre 2017 et entendu de nombreuses personnalités qualifiées. Le rapport *Réseaux sociaux et Santé* rend compte de ses travaux.

Il décrit les principaux usages des réseaux sociaux de santé et les catégories de données générées par les patients sur ces réseaux : échanges au sein de communautés de patients sur leur pathologie et leur prise en charge, débats sur des enjeux de santé publique, partage par les utilisateurs d'informations sur leur santé avec leurs «amis» ou «abonnés» n'appartenant pas à une communauté de patients. Il précise les principaux lieux de partage (Facebook et Twitter). (I)

Il décrit également les modalités actuelles d'accès par des tiers aux données échangées sur les principaux réseaux (Facebook et Twitter), les données auxquelles il est possible d'accéder comme les données pour lesquelles aucun accès n'est possible. (II)

Il recense ensuite les principales utilisations actuelles des données de santé partagées sur les réseaux : la communication ciblée (campagne de santé publique, recrutement de participants à une étude ou un programme de recherche), leurs finalités (suivi épidémiologique, pharmacovigilance, analyse des parcours de soins) et les méthodologies utilisées. (III)

À l'issue de cette étude, le groupe de travail a conclu que :

• les réseaux sociaux sont devenus une source complémentaire de données de vie réelle[2] qui doit être prise en considération, notamment par les pouvoirs publics dans le cadre de leurs missions de veille sanitaire et de prévention, comme le font déjà des autorités étrangères[3] ;

• ces données, précieuses pour les études populationnelles, ne permettent toutefois pas à elles-seules de contextualiser un évènement à l'échelle individuelle. Pour ces usages, une approche mixte et des appariements des données issues des réseaux avec des données médicalisées doivent être envisagés ;

• la multiplication des études de données textuelles des réseaux et de données textuelles en général, influera sur les pratiques de recherche qui évolueront vers une plus grande prise en compte des données de vie réelle ;

1. Ce groupe était composé d'un spécialiste de l'étude des données de santé sur les réseaux sociaux (Adel Mebarki – Kap Code), de spécialistes de la data science appliquée à l'épidémiologie et la santé publique (Francisco Orchard – Epiconcept), d'un professionnel de l'analyse des données médico-économiques (Patrick Olivier – IVBAR France), d'un représentant de la direction digitale d'un laboratoire pharmaceutique (Lina Autelitano – Pierre Fabre) et d'un avocat en droit des nouvelles technologies et des données personnelles appliquées à la santé (Caroline Henry – Pons & Carrère).

2. *«On désigne sous le terme «données de vie réelle», ou «données de vraie vie», des données qui sont sans intervention sur les modalités usuelles de prise en charge des malades et qui ne sont pas collectées dans le cadre expérimental (le cadre notamment des essais randomisés contrôlés, ECR), mais qui sont générées à l'occasion des soins réalisés en routine pour un patient, et qui reflètent donc a priori la pratique courante. De telles données peuvent provenir de multiples sources : elles peuvent être extraites de dossiers informatisés de patients, ou constituer un sous-produit des informations utilisées pour e remboursement des soins ; elles peuvent être collectées de manière spécifique, par exemple dans le cadre de procédures de pharmacovigilance, ou pour constituer des registres ou des cohortes, ou plus ponctuellement dans le cadre d'études ad hoc ; elles peuvent également provenir du web, de réseaux sociaux, des objets connectés, etc.»*
Bégaud, D. Polton, F. von Lennep, Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé, Rapport réalisé à la demande de Madame La Ministre de la santé Marisol Touraine, mai 2017, https://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf, page 4

3. La FDA (Food and Drug Administration) qui nouait dès 2015 un partenariat avec Google http://www.iracm.com/2015/07/google-et-la-fda-creent-un-partenariat-afin-dutiliser-les-donnees-du-moteur-de-recherche-pour-identifier-les-effets-indesirables-des-medicaments/ ; la Société américaine de cardiologie ou les Nations Unies dans le cadre du programme Global Pulse https://www.unglobalpulse.org/projects

• ces évolutions ne se feront pas sans :

- une explication du raisonnement suivi et une évaluation de la performance des algorithmes (précision, compréhension, reproductibilité, etc.) ;

- une réflexion approfondie sur la sécurité des données partagées sur les réseaux. Cette réflexion doit être celle des réseaux eux-mêmes, des pouvoirs publics mais aussi des personnes concernées. Une juste information sur les données « détenues » par les réseaux sociaux et la sécurité qui peut raisonnablement être attendue doit permettre à chacun de prendre une décision éclairée sur les espaces au sein desquels il souhaite effectivement partager ses données.

Le groupe de travail a formulé **quatre propositions** pour une exploitation responsable des données de santé générées par les patients sur les réseaux sociaux, dans l'intérêt des patients et le respect de leurs droits :

1. Favoriser l'exercice effectif du droit à la **portabilité** des données de santé générées sur les réseaux sociaux pour permettre l'exploitation de **toutes** les données partagées, **à l'initiative et sous le contrôle des utilisateurs :**

- Les utilisateurs devraient pouvoir « récupérer » leurs données organisées selon les thématiques qui ont généré leur activité (dont la santé) et non selon les catégories d'activités propres aux réseaux sociaux (publications, commentaires, mentions « J'aime » etc.) et les données de l'activité d'un compte devraient pouvoir être partiellement portables, thématique par thématique ;

- des fonctionnalités devraient être mises à la disposition des utilisateurs pour leur permettre de transférer directement leurs données à un porteur de projet de recherche, via par exemple des outils simples comme des appels au partage de données, sur le modèle des outils existants d'appel aux dons[4].

2. **Sensibiliser** les citoyens à l'existence d'API et à leur fonctionnement. Les **informer, précisément et de manière simple et adaptée**, des conditions d'utilisation et des catégories d'utilisateurs des API des réseaux sociaux qu'ils utilisent, pour permettre une exploitation transparente, respectueuse des droits des utilisateurs des réseaux et **sécurisée** pour les exploitants de données.

3. Permettre un accès **simplifié et gratuit** aux bases de données comprenant les **données rendues publiques sur les réseaux sociaux par leurs utilisateurs,** pour les **acteurs de la recherche publique** et les porteurs de projets de recherche scientifique financés par les pouvoirs publics ou commandés par les pouvoirs publics dans l'exercice de leurs missions de service public. À cette fin, l'exception légale dite de *data mining*[5] pourrait être utilement élargie.

4. Développer des **partenariats entre les pouvoirs publics et les réseaux sociaux** et, dans ce cadre, mettre à la disposition des pouvoirs publics des **modules de ciblage avancé** pour permettre une **diffusion efficace des campagnes publiques de prévention** ou le **recrutement de volontaires pour les projets de recherche publique** ou les projets conduits pour le compte des pouvoirs publics dans l'exercice de leurs missions de service public.

4. https://fr-fr.facebook.com/help/990087377765844?helpref=faq_content

5. Article L.342-3 du code de la propriété intellectuelle

TABLE DES MATIÈRES

INTRODUCTION p. 8

PARTIE I. LES DONNÉES RELATIVES À LA SANTE GÉNÉRÉES SUR LES RÉSEAUX SOCIAUX p. 12

A. LES PARTAGES DE DONNÉES DE SANTÉ SUR LES RÉSEAUX : UN CONSTAT DES ASSOCIATIONS DE PATIENTS p. 12

B. LE PARTAGE DES DONNÉES DE SANTÉ SUR LES RÉSEAUX : UN PHÉNOMÈNE DE GRANDE AMPLEUR p. 13

C. LES LIEUX DE PARTAGE p. 15

PARTIE II. L'ACCÈS AUX DONNÉES DE RÉSEAUX SOCIAUX p. 20

A. LE MÉCANISME D'ACCÈS AUX DONNÉES DES UTILISATEURS DE FACEBOOK ET INSTAGRAM p. 20

1. Les API et autorisations d'accès technique p. 20

2. L'autorisation ou l'information des personnes concernées p. 22

3. Les observations du groupe de travail p. 23

B. L'ACCÈS AUX DONNÉES DES UTILISATEURS DE TWITTER p. 25

1. Twitter et les données « publiques » p. 25

2. Les API de Twitter p. 26

3. Les observations du groupe de travail p. 27

PARTIE III. USAGES DES DONNÉES DES RÉSEAUX SOCIAUX EN SANTÉ p. 34

A. LA COMMUNICATION CIBLÉE p. 34

1. Les campagnes d'information de santé publique p. 34

2. La proposition de biens et services de santé à destination des patients p. 35

3. Le recrutement de participants pour un mouvement/ une étude/un programme de santé p. 35

4. Les observations du groupe de travail p. 36

B. OBSERVATION DE POPULATIONS / RECHERCHE p. 38

1. Les projets de recherche conduits sur les données de réseaux p. 38

2. Méthodologie d'analyse des données des réseaux p. 41

C. PERSPECTIVES p. 43

4 PROPOSITIONS Pour une exploitation responsable des données de santé générées par les patients sur les réseaux sociaux dans l'intérêt des patients et le respect de leurs droits p. 44

ANNEXE I : LES OUTILS D'ANALYSE DES DONNÉES DES RÉSEAUX p. 46

ANNEXE II : PERSONNES ENTENDUES ET REMERCIEMENTS p. 50

QUELQUES REPÈRES BIBLIOGRAPHIQUES p. 52

INTRODUCTION

Une multitude de données de santé, issues de sources diverses, est produite au sujet d'un patient, qu'il soit traité ponctuellement ou suivi pour une pathologie chronique. Mais le patient génère, parallèlement à ces données cliniques ou médico-administratives, d'autres données, lorsqu'il s'exprime à propos de sa santé sur des forums ou des réseaux sociaux.

Les entretiens que nous avons menés nous ont conduits à un premier constat : les patients s'expriment plus spontanément à propos de leur pathologie sur les réseaux sociaux qu'auprès des professionnels de santé. Ils y décrivent leur humeur, évoquent leurs conditions de vie, l'expérience qu'ils font du traitement, les bénéfices qu'ils en retirent, l'inconfort qu'ils constatent ou les effets indésirables qu'ils ressentent. La parole partagée sur les réseaux est récurrente et librement exprimée, c'est-à-dire non déterminée par les questions posées.

L'information livrée dans ce contexte est difficilement accessible aux professionnels de santé comme aux chercheurs pour deux raisons au moins. Dans le cadre du suivi médical, l'échange est plus épisodique et certains évènements, survenus entre deux consultations, sont oubliés. La relation d'autorité, même bienveillante, exerce également sur le patient un puissant effet inhibiteur. Ainsi, celui-ci se contentera souvent de répondre aux interrogations formulées par les professionnels en charge de son suivi ou conduisant des recherches et certains éléments jugés anodins voire gênants, seront tus.

Il existe donc sur les réseaux sociaux des données de « vraie vie »[6] générées par les patients et susceptibles, à l'échelle individuelle, de compléter utilement le dossier médical d'un patient, à l'échelle collective, de faire progresser la connaissance (I.). Mais ces données sont-elles pour autant exploitables ? Sont-elles accessibles, sous quelles conditions et avec quels objectifs ? (II.)

En réalité, ces données sont déjà utilisées, tel que nous allons le montrer dans ce rapport (III.).

Sur la base des premiers cas d'usage que nous avons recensés et exposés, nous avons tenté d'imaginer les perspectives d'avenir qu'ouvre ce nouveau champ. Nous serons ainsi amenés à formuler des propositions pour que les données issues des réseaux sociaux puissent être exploitées de manière responsable, dans l'intérêt individuel et collectif des patients et dans le respect de leurs droits.

6. Voir note 2.

PARTIE I.

LES DONNÉES RELATIVES À LA SANTÉ GÉNÉRÉES SUR LES RÉSEAUX SOCIAUX

LES DONNÉES RELATIVES À LA SANTÉ GÉNÉRÉES SUR LES RÉSEAUX SOCIAUX

A. LES PARTAGES DE DONNÉES DE SANTÉ SUR LES RÉSEAUX : UN CONSTAT DES ASSOCIATIONS DE PATIENTS

Convaincu que l'exploitation des données de santé ne peut se faire **que dans l'intérêt des patients,** mais aussi avec les patients, le Healthcare Data Institute a initié, au début de l'année 2017, une réflexion sur les patients et leurs données. Au mois d'avril 2017, une réunion de lancement était organisée avec des patients et représentants d'associations de patients. Ces premiers échanges ont montré que le partage de certaines données avec les professionnels en charge du suivi des patients ou la contribution à des projets de recherche pouvait susciter certaines appréhensions ou interrogations.

Le partage de ces mêmes données avec une communauté d'amis, de famille ou de patients sur les réseaux sociaux était en revanche présenté comme une pratique fréquente et spontanée, motivée par le soutien et le réconfort d'un «like» ou d'un commentaire bienveillant, le besoin de s'exprimer à l'instant présent et sur le thème choisi par le patient ou l'envie de partager son expérience ou de bénéficier de conseils de ses pairs.

Partant de ce constat et de l'intuition que la masse de données partagées sur les réseaux pourrait contribuer à une meilleure prise en charge des patients et au développement de la connaissance, le Healthcare Data Institute a décidé de structurer un groupe de travail pluridisciplinaire pour analyser les enjeux et perspectives d'usage des données générées sur les réseaux sociaux de santé. Ce groupe était composé d'un spécialiste de l'étude des données de santé sur les réseaux sociaux[7], de spécialistes de la data science appliquée à l'épidémiologie et la santé publique[8], d'un professionnel de l'analyse des données médico-économiques[9], d'un représentant de la direction digitale d'un laboratoire pharmaceutique[10] et d'un avocat en droit des nouvelles technologies et des données personnelles appliquées à la santé[11].

7. Adel Mebarki – Kap Code

B. LE PARTAGE DES DONNÉES DE SANTÉ SUR LES RÉSEAUX : UN PHÉNOMÈNE DE GRANDE AMPLEUR

Le groupe de travail a débuté ses travaux en septembre 2017.

Après un premier partage d'expériences et de réflexions, le groupe a organisé plusieurs cycles d'auditions ou d'entretiens qui lui ont permis d'entendre, au cours du premier semestre 2018, des représentants ou fondateurs d'associations de patients, des professionnels des réseaux sociaux, des chercheurs conduisant des projets de recherche sur la base de données générées sur les réseaux, des professionnels proposant des analyses de données de réseaux, des créateurs de réseaux sociaux dédiés aux patients ou d'une plateforme de recherche collaborative.

Dans le même temps, le groupe a entrepris un travail de recherche documentaire, scientifique et juridique.

Un appel à contributions des membres du Healthcare Data Institute a été lancé en avril 2018.

Une première version de travail du rapport a été communiquée aux personnes rencontrées.

Ces travaux confirmaient le phénomène décrit par les patients et associations de patients.

Au fil des années, accompagnant la tendance sociétale générale, les patients se sont progressivement exprimés sur Internet. Comme pour les

8. Francisco Orchard – Epiconcept

9. Patrick Olivier – IVBAR France

10. Lina Autelitano – Pierre Fabre

11. Caroline Henry – Pons & Carrère

autres utilisateurs, leur expression est devenue de plus en plus quotidienne, de plus en plus ciblée sur les détails de leur vie privée (leur suivi médical et le quotidien de leur pathologie) et de moins en moins anonyme. Tout a commencé avec les premiers forums santé comme Doctissimo, suivis par les blogs de patients comme celui d'Yvanie Caillé sur les maladies rénales puis enfin les groupes Facebook, fils Twitter (#DOC pour le diabète par exemple) ou autres trends Instagram. Les utilisateurs de réseaux sociaux s'y expriment aujourd'hui sur leur santé, comme sur tout autre aspect de leur quotidien, et les usages se sont progressivement diversifiés. L'expression s'est ainsi déplacée de supports thématiques vers des supports plus généralistes avec une segmentation opérée par les utilisateurs eux-mêmes.

Si le phénomène est toutefois difficile à quantifier, quelques chiffres tendent néanmoins à en démontrer l'ampleur. Selon le sondage Opinon Way pour le Festival de la communication santé et Ramsay Générale de Santé « Réseaux sociaux santé dans la prévention et l'éducation thérapeutique des Français[12] » par exemple, 89% des Français souffrant d'une pathologie percevraient un bénéfice à utiliser les réseaux sociaux dans le cadre de leur santé et un quart des Français jugerait les réseaux utiles pour partager des informations avec d'autres personnes souffrant des mêmes problèmes de santé.

Le groupe a recensé trois raisons principales motivant aujourd'hui les utilisateurs à s'exprimer sur leur santé ou la santé, sur les réseaux sociaux :

• **Participer au débat public sur la santé :** L'utilisateur citoyen prend part aux débats d'actualité sur les réseaux sociaux. Il affiche – parfois avec véhémence – ses positions sur différents sujets comme la vaccination, l'IVG ou les réformes du gouvernement. En ce qui concerne la vaccination par exemple, les échanges sur les réseaux sociaux représentent environ 250 000 partages d'articles sur Facebook, 46 000 tweets et 1 500 discussions chaque mois sur l'année 2018 (source Antidox).

• **Échanger au sein d'une communauté de patients :** Les réseaux sociaux permettent aux utilisateurs de se regrouper et de construire des communautés articulées autour d'une même problématique de santé, comme une maladie chronique commune. Ces communautés, souvent très actives, sont l'espace idéal pour des échanges d'expériences entre pairs, des soutiens ou des réconforts face au fardeau de la maladie. On peut également y trouver des réponses aux questions visant à améliorer le quotidien des malades.

Des leaders d'opinion animent ces communautés. Ils peuvent représenter leurs membres et plaider pour leurs intérêts communs. C'est le cas par exemple pour les patients atteints d'hypoparathyroïdie qui partagent leur expérience dans plus de 50% des cas sur le forum «Vivre sans Thyroïde».

• **Communiquer sur son état de santé :** Partageant son quotidien, l'utilisateur évoque sa santé, de plusieurs façons. Son témoignage va d'un **statut Facebook** «cloué au lit avec la grippe» à une publication liée à un **objet connecté** comme le partage du suivi de l'évolution de son poids. A l'extrême, on trouve des **patients-utilisateurs** affichant fièrement et publiquement leur combat contre la maladie. Plusieurs mannequins se sont ainsi mises en scène sur Instagram, comme Gianni Marshall exposant les taches apparues sur son visage suite à un accident ischémique transitoire.

C'est précisément sur cette communication directe et spontanée qui génère de nouvelles données que se concentreront ces travaux.

C. LES ESPACES DE PARTAGE

Les travaux du groupe ont permis de distinguer trois sous-catégories d'espaces d'échange et de partage:

- **les réseaux sociaux généralistes :** Facebook, Twitter, Instagram, YouTube, Snapchat ;
- **les forums publics et blogs :** Doctissimo, Renaloo, Sante AZ, Onmeda, AuFeminin etc.;
- **les communautés de patients :** Carenity, Entrepatients, PatientsLikeMe, Seronet, Respire, etc.

Nous disposons aujourd'hui d'assez peu de recul sur les communautés de patients en France, encore jeunes. En revanche, l'expérience montre que

12. https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/marketing/sante/opinionway-pour-le-festival-de-la-communication-sante-francais-et-reseaux-sociaux-sante-novembre-2016.html

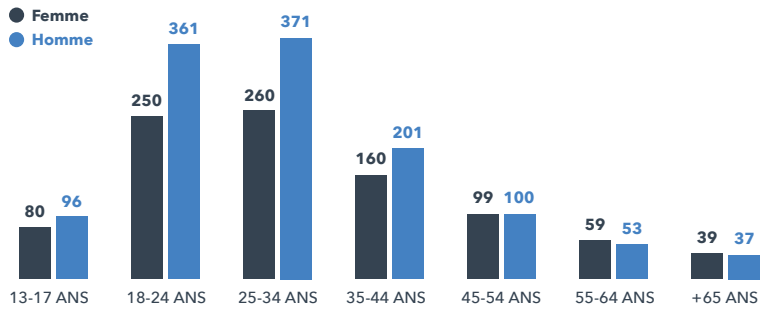
les utilisateurs recherchant des informations relatives à leur pathologie interviendront plutôt, souvent **de façon anonyme**, sur des forums dédiés[13].

Des utilisateurs plus jeunes privilégieront les groupes Facebook ou Twitter où ils s'identifient pour exprimer leur perception de la maladie ou afficher publiquement leurs positions et revendications.

D'une manière générale, les réseaux sociaux généralistes seront utilisés par leurs membres pour partager des données relatives à leur santé, parmi d'autres données de leur vie privée. Le phénomène n'est pas anecdotique puisque Facebook, par exemple, compte aujourd'hui 34 millions d'utilisateurs actifs mensuels en France. Selon les chiffres du rapport 2018 publié conjointement par Hootsuite et We Are Social, 67% des Français seraient actifs sur Facebook, avec un élargissement général des tranches d'âges connectées au réseau. Après Facebook Messenger et YouTube, 24% des Français utiliseraient aussi Twitter.

Nous avons donc fait le choix, dans ce rapport, de concentrer l'étude sur les réseaux généralistes, et plus particulièrement sur Facebook et Twitter.

PROFILE OF FACEBOOK USERS
BREAKDOWN OF FACEBOOK'S GLOBAL USERS BY AGE AND GENDER, IN MILLIONS



Janvier 2018

13. Sondage Opinon Way pour le Festival de la communication santé et Ramsay Générale de Santé «Réseaux sociaux santé dans la prévention et l'éducation thérapeutique des Français», publié en novembre 2016 https://www.opinion-way.com/fr/sondage-d-opinion/sondages-publies/marketing/sante/opinionway-pour-le-festival-de-la-communication-sante-francais-et-reseaux-sociaux-sante-novembre-2016.html

PARTIE II.

L'ACCÈS AUX DONNÉES DE RÉSEAUX SOCIAUX

L'ACCÈS AUX DONNÉES DE RÉSEAUX SOCIAUX

A. LE MÉCANISME D'ACCÈS AUX DONNÉES DES UTILISATEURS DE FACEBOOK ET INSTAGRAM

1. LES API ET AUTORISATIONS D'ACCÈS TECHNIQUE

Les réseaux Facebook et Instagram permettent à des tiers d'accéder aux données des utilisateurs grâce à des API (application program interface ou interface de programmation d'application)[14]. Cet accès est prévu par la politique de confidentialité de ces plateformes[15] et peut être techniquement schématisé de la manière suivante :

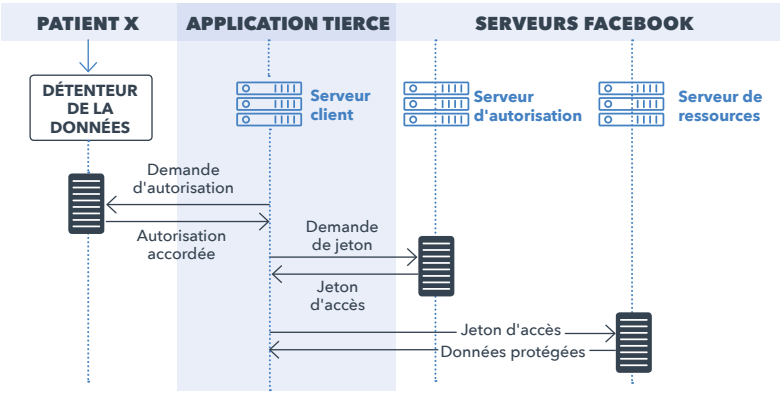


Schéma d'accès aux données de réseaux via les API

14. *« L'API peut être résumée à une solution informatique qui permet à des applications de communiquer entre elles et de s'échanger mutuellement des services ou des données. Il s'agit en réalité d'un ensemble de fonctions qui facilitent, via un langage de programmation, l'accès aux services d'une application ».* https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203559-api-application-programming-interface-definition-traduction/

15. https://www.facebook.com/about/privacy/update/printable

Pour mettre en œuvre cette technologie, Facebook met à la disposition des professionnels de nombreux outils (**login via Facebook** et boutons sociaux par exemple).

Les données mises à disposition des tiers via les API sont dites *« Propriétés étendues du profil »*. Elles sont classées **selon les catégories d'autorisations d'accès technique**[16] **accordées** par Facebook ou Instagram après **revue de l'application tierce**. Les principales autorisations d'accès peuvent être présentées de la manière suivante :

DONNÉES UTILISATEURS

- **URL du profil**
- **Tranche d'âge**
- **Sexe**
- **Date d'anniversaire** *(selon les autorisations)*
- **Ville natale**
- **Ville actuelle**

DONNÉES SOCIALES

- **Likes** *(Pages/Films/Séries etc.)*
- **Lieux visités**
- **Événements** *(acceptés/organisés)*
- **Amis**
- **Photos/Vidéos** *(selon les autorisations)*
- **Statuts** *(selon les autorisations)*

DONNÉES UTILISATEURS

- **URL du profil**
- **Tranche d'âge**
- **Sexe**
- **Date d'anniversaire** *(selon les autorisations)*
- **Ville actuelle**

DONNÉES SOCIALES

- **Médias partagés** *(photos/vidéos)*
- **Commentaires**
- **Likes**
- **Abonnés/Abonnements**

D'autres données étaient rendues accessibles auparavant. Ces accès ont été fermés le 4 avril 2018[17], suite à l'affaire dite *Cambridge Analytica*. Des études semblent toutefois montrer que même après la mise en œuvre de ces modifications, les API Facebook sont susceptibles de permettre un accès plus large aux données des utilisateurs[18].

16. https://developers.facebook.com/docs/facebook-login/permissions/?translation##reference-default_fields

17. user_website, user_work_history, read_custom_friendlists, user_about_me, user_actions.books, user_actions.fitness,user_actions.music,user_actions.news, user_actions.video, user_actions:{app_namespace}, user_education_history, user_games_activity, user_groups, user_relationship_details, user_relationships, user_religion_politics, user_status.

18. https://www.wired.com/story/security-risks-of-logging-in-with-facebook/ "The security risks of logging in with Facebook"

2. L'AUTORISATION OU L'INFORMATION DES PERSONNES CONCERNÉES

Une fois l'autorisation de Facebook obtenue, le tiers doit obtenir **le consentement de l'utilisateur** pour accéder à ses données ou l'informer de cet accès. Cette demande ou information se présente ainsi :

L'étendue du partage peut être modifiée par l'utilisateur qui ne peut toutefois refuser le partage de ses données de profil publiques.

Les données dites *«données de profil étendu»* des utilisateurs de Facebook ne peuvent donc être partagées avec des tiers qu'à la double condition que Facebook et l'utilisateur concerné l'y autorisent.

Cela signifie que l'accès aux données d'un utilisateur peut être refusé alors même que celui-ci serait disposé à partager ses «données Facebook» avec des tiers, dans le cadre d'un projet de recherche ou d'un dispositif de santé publique.

Dans le même temps, les conditions d'utilisation des données de Facebook prévoient que ces mêmes données peuvent être utilisées par la plateforme et ses «partenaires» pour *«faire de la recherche et innover pour le bien-être social»* : *«Nous utilisons les informations à notre disposition (notamment celles de partenaires de recherche avec lesquels nous collaborons) pour orienter et appuyer la recherche et l'innovation sur des sujets de bien-être social général, d'avancement technologique, d'intérêt public, **de santé et de bien-être**[19].»*

3. LES OBSERVATIONS DU GROUPE DE TRAVAIL

Ce système d'exploitation des données des profils «privés» des utilisateurs pose pour le groupe de travail des difficultés pour deux raisons :

• la licéité du traitement des données résultant des échanges des utilisateurs avec leurs «amis» (commentaires, statuts, «likes») : Si la question reste en partie incertaine, la jurisprudence française tend à reconnaître à ces données le caractère de correspondances privées, prenant néanmoins en considération, au cas par cas, les paramétrages de confidentialité et le nombre d'«amis» de l'utilisateur[20].

• la légitimité du contrôle préalable par Facebook des applications pouvant accéder aux informations des utilisateurs : Les réseaux disposent assurément de droits de propriété intellectuelle sur les sites qu'ils exploitent et sur la technologie qui en permet le fonctionnement. Cela implique-t-il pour autant leur accord préalable pour accéder aux données des utilisateurs, notamment à des fins de recherche et d'étude ou de mise en œuvre de dispositifs de santé publique ?

Le groupe de travail est d'avis que l'effectivité du droit à la portabilité, prévu par le Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (RGPD), pourrait favoriser dans ce contexte une exploitation responsable des données des réseaux sociaux.

---

19. Politique d'utilisation des données de Facebook : Comment utilisons-nous ces informations ? https://research.fb.com

20. Cass Civ 1ère 10 avril 2013, Bul. Civ I n°70 ; Depuis la loi pour une République Numérique, anticipant sur le Règlement e-privacy, les réseaux sociaux sont astreints, de manière expresse, au respect du secret des correspondances. Ce secret leur impose de ne pas traiter les données contenues dans ces correspondances si ce n'est avec le consentement spécifique des utilisateurs concernés et pour des finalités restreintes : à des fins publicitaires, statistiques ou d'amélioration du service apporté à l'utilisateur. La finalité de recherche scientifique est-elle ici aussi nécessairement compatible ? https://www.legifrance.gouv.fr/affichJuriJudi.do?idTexte=JURITEXT000027303638

Le droit à la portabilité des données, prévu par l'article 20 du Règlement, permet aux personnes concernées – ici les utilisateurs – de recevoir du responsable du traitement les données qui les concernent et qu'elles lui ont fournis. Cette transmission doit se faire dans un format qui permet la réutilisation de ces informations. L'article 20 prévoit également au bénéfice des personnes concernées, la faculté de transmettre leurs données à un tiers, responsable d'un second traitement des données. Elles peuvent même demander que ces données soient directement transmises par le premier responsable de traitement, au second.

Dans le contexte des données de santé partagées sur les réseaux sociaux, l'exercice du droit à la portabilité des données permettrait un contrôle par les patients des usages faits de leurs données, en même temps qu'une exploitation de ces données par les professionnels en charge du suivi des patients ou par des chercheurs, et répondrait aux préoccupations de licéité et de légitimité susmentionnées.

Toutefois, le droit à la portabilité des données suppose que les données puissent être **effectivement réutilisées par les personnes concernées.** Cette possibilité de réutilisation des données reste aujourd'hui très limitée, puisque les données restituées aux utilisateurs sont organisées selon les catégories d'activités spécifiques aux réseaux sociaux (commentaires, «likes», statuts, tweets, retweets) et non selon le sujet ou l'objet de ces données.

Pour favoriser l'effectivité du droit à la portabilité des données de santé partagées sur les réseaux sociaux, le groupe considère que :

• la question du format de restitution des données des utilisateurs devrait aussi être envisagée sous l'angle de la structuration/de la sélection des données des utilisateurs :

  - les utilisateurs devraient pouvoir «récupérer» leurs données organisées selon les thématiques qui ont généré leur activité (dont la santé) ;

  - les données de l'activité d'un compte devraient pouvoir être partiellement portables, thématique par thématique.

• des fonctionnalités devraient par ailleurs être mises à la disposition directe des utilisateurs pour leur permettre de transférer directement leurs données à un porteur de projet de recherche, via par exemple des outils simples comme des appels au partage de données, qui pourraient être similaires aux outils existants d'appel aux dons[21].

B. L'ACCÈS AUX DONNÉES DES UTILISATEURS DE TWITTER

1. TWITTER ET LES DONNÉES «PUBLIQUES»

La problématique de l'accès aux données publiées sur les pages du réseau Twitter est un peu différente, puisque les informations sont voulues publiques par les utilisateurs, sauf exception, c'est-à-dire publiées sans restriction d'accès pour les tiers.

La politique de confidentialité[22] de Twitter compte parmi les données «publiques» :

• les informations de profil,

• le fuseau horaire et la langue utilisée par la personne concernée,

• la date de création de son compte,

• les tweets et certaines informations sur les tweets, telles que la date, l'heure, l'application et la version de Twitter à partir desquelles l'utilisateur a tweeté,

• le lieu d'émission des tweets ou la localisation précisée sur le profil Twitter,

• les listes créées par les utilisateurs, les personnes «suivies» et qui suivent l'utilisateur,

• les tweets «aimés» ou «retweetés», et

• les informations postées au sujet de l'utilisateur par d'autres utilisateurs (par exemple, lorsque la personne est «taggée» sur une photographie ou mentionnée dans un tweet).

21. https://fr-fr.facebook.com/help/990087377765844?helpref=faq_content

22. https://twitter.com/fr/privacy

Cette «publicité» est «l'ADN» de Twitter, comme le rappelle la plateforme : *«Twitter est conçu pour diffuser instantanément et en masse des informations que vous partagez publiquement à travers nos services»*. Elle est en grande partie le fondement de la politique d'utilisation des données de Twitter[23]. La publicité des informations ou contenus implique pour Twitter une «licence» gratuite, un consentement des utilisateurs à toutes formes d'exploitation par les tiers de leurs «informations publiques», quel que soit le type d'information partagée[24]. Ce mécanisme est explicitement présenté dans les conditions générales d'utilisation du site[25].

2. LES API DE TWITTER

Pour permettre cette exploitation des «informations publiques», la plateforme a implémenté des API dont elle conditionne aussi l'utilisation à une autorisation préalable et au paiement d'un prix plus ou moins conséquent selon les performances de l'API. Ces API sont nombreuses. Les API qui permettent un accès à la base des Tweets peuvent être schématisées ainsi :

APIs	DESCRIPTION	EXHAUSTIVITÉ DES DONNÉES	HISTORIQUE ACCESSIBLE	TARIFICATION	LIMITES
Search Standard	Échantillon aléatoire de tweets publics à partir d'un ou plusieurs mots clés	Incomplète	7 jours	Gratuite	Échantillon aléatoire
Search Tweet Premium	L'ensemble des tweets postés à partir d'un ou plusieurs mots clés dans la limite de 5 millions de tweets	Complète	30 jours	De 149$ à 2 499$	500 tweets maximum par requête/Nombre de requêtes limité
Search Tweet Premium	L'ensemble des tweets postés à partir d'un ou plusieurs mots clés dans la limite de 5 millions de tweets	Complète	L'ensemble des tweets depuis 2006	De 99$ à 1 899$	500 tweets maximum par requête/Nombre de requêtes limité
Entreprise Premium	L'ensemble des tweets postés à partir d'un ou plusieurs mots clés sans limite de nombre de tweets	Complète	30 jours	Sur devis	Coût important/Limite temporelle
Entreprise Premium	L'ensemble des tweets postés à partir d'un ou plusieurs mots clés sans limite de nombre de tweets	Complète	L'ensemble des tweets depuis 2006	Sur devis	Coût important

23. qui couvre également les données de Periscope©

24. Définis comme : *«Les informations, textes, liens, graphiques, photos, sons, vidéos ou autres éléments ou combinaison d'éléments téléchargés à partir des Services ou apparaissant sur ceux-ci (collectivement dénommés le "Contenu")»*, https://twitter.com/fr/tos

25. https://twitter.com/fr/tos «Vos droits et Concession de droits sur le Contenu»

3. LES OBSERVATIONS DU GROUPE DE TRAVAIL

La logique de licence d'utilisation consentie par les utilisateurs sur leurs contenus est cohérente avec la proposition de services qui leur est faite puisque :

- ceux-ci veulent en principe que leurs communications, Tweets ou autres activités sur leur compte soient publics, et

- Twitter ne peut exercer concrètement un contrôle sur l'exploitation des Tweets par les tiers.

Toutefois, elle ne permet pas à notre sens de justifier la réutilisation des données des utilisateurs via les API de recherche. Les utilisateurs ne peuvent par avance consentir aux traitements qui seraient effectués par des tiers dont ils ignorent tout. La publicité de leurs contenus, voulue ou acceptée par les utilisateurs, ne peut équivaloir, comme le suggèrent les CGU de Twitter, à un consentement éclairé de ceux-ci à tout traitement de leurs données personnelles qui pourrait être fait par des tiers, notamment via les API de recherche.

Ces API, du reste, ne contiennent pas de fonctionnalité permettant d'informer directement les utilisateurs du traitement qui va être fait de leurs données, ni même de recueillir le cas échéant leur accord pour y accéder.

Cette publicité peut néanmoins rendre licite le traitement de données de santé publiées sur Twitter[26], à charge pour le tiers accédant à ces données de se mettre en conformité avec la réglementation dans le cadre du traitement dont il est responsable. Le nombre des utilisateurs concernés, la logistique et les coûts nécessaires pourraient également justifier l'absence d'information des utilisateurs concernant le traitement[27].

Reste que le traitement licite des données suppose que la collecte initiale ait été régulièrement faite et notamment, que les utilisateurs de Tweeter aient été dûment informés des finalités et catégories de destinataires de leurs données. Et sur ce point, les documents contractuels proposés en ligne sont évasifs. Les API sont rapidement présentées dans

26. Article 9 e) RGPD

27. Article 14.5. b) RGPD

la section consacrée aux informations que les utilisateurs partagent avec Twitter, et non pas dans la section relative aux partages faits par Twitter des informations des utilisateurs. Aucune précision n'est apportée sur les catégories de tiers qui pourraient ainsi avoir accès aux données et les quelques exemples de dispositifs de veille sanitaire ayant utilisé les données des utilisateurs ne sont pas immédiatement accessibles[28].

Le groupe a par ailleurs constaté que Twitter, aux termes des conditions générales d'utilisation, limite les possibilités de crawling[29] et prohibe le scraping[30] sans son accord préalable. Cette contrainte impose donc aux tiers de passer par Twitter et d'utiliser ses API, dans leur version payante, notamment pour la recherche qui nécessite des échantillons de données significatifs.

Cela revient à valoriser, non seulement la technologie développée par le réseau, mais également les données des utilisateurs, données que les utilisateurs ont voulu publiques, accessibles et utilisables par le plus grand nombre, suivant la logique de Twitter.

28. Par décision en date du 7 août 2018, le Tribunal de grande instance de Paris vient de déclarer illicites 265 clauses actuelles ou anciennes des conditions d'utilisation, de la politique de confidentialité et des règles sur Twitter, parmi lesquelles semble-t-il *«les clauses autorisant à copier, modifier, adapter, vendre les contenus passés et futurs des utilisateurs, y compris ceux protégés par le droit de la propriété intellectuelle, à tout bénéficiaire, sur tout support, sans autorisation préalable»* https://www.legalis.net/actualite/ladhesion-a-twitter-est-un-contrat-de-consommation/ ; les autorités européennes demandent également à Facebook et Twitter de revoir les clauses de leurs conditions générales relatives à l'utilisation des données des utilisateurs https://www.usinenouvelle.com/article/l-ue-s-impatiente-et-menace-facebook-et-twitter-de-sanctions.N744214 ; https://www.theverge.com/2018/9/19/17880348/facebook-european-union-fines-sanctions-terms-of-service-changes

29. Le crawling consiste à utiliser un crawler, ou spider. Un crawler est un robot d'indexation. *«Il s'agit d'un logiciel qui a pour principale mission d'explorer le Web afin d'analyser les contenus ainsi explorés. Le crawler parcourt donc, de façon autonome et automatique, les différents sites et pages Internet à la recherche de contenus bien précis ou d'éventuelles mises à jour. Derrière cette activité se cache une autre mission : celle d'indexer les pages Web en fonction de la qualité des contenus et, ainsi, aider les moteurs de recherche à classer les pages Internet dans l'affichage des résultats.»* https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203499-crawler-definition-traduction-et-acteurs/

30. CGU article 4 : *«Lorsque vous accédez aux Services ou que vous les utilisez, vous vous interdisez de faire ce qui suit : (…) ((iii) rechercher les Services, y accéder, ou tenter de les rechercher ou d'y accéder par tout moyen (automatisé ou non) autre que les interfaces actuellement disponibles, développées et fournies par Twitter (sous réserve de respecter les conditions générales en vigueur), à moins que vous n'y ayez été expressément autorisé(e) aux termes d'un accord séparé avec Twitter (NOTE : le crawling [indexation systématique] des Services est autorisé si cela est fait conformément aux dispositions du fichier robots. txt ; par contre, le scraping [extraction à des fins d'exploitation] des Services sans l'accord préalable de Twitter est expressément interdit)»*

Il y a donc un frein potentiel à l'exploitation des données partagées sur ce réseau dû à la nécessité d'être autorisé à utiliser les API et surtout, de régler le coût de leur utilisation qui peut être important, notamment pour un projet de recherche public ou conduit à l'initiative des pouvoirs publics.

Cette «barrière à l'entrée» est assez contradictoire avec la raison d'être affichée par le réseau. La valorisation des données des utilisateurs va également au-delà du modèle économique publicitaire du réseau que les conditions d'utilisation du réseau rappellent aux utilisateurs : *«En contrepartie du droit à accéder et à utiliser les Services qui vous est consenti par Twitter, vous acceptez que Twitter et ses prestataires et partenaires tiers puissent placer des publicités sur les Services, ou en relation avec l'affichage du Contenu ou des informations provenant des Services et soumis par vous ou par d'autres[31].»*

Pour répondre à ces problématiques, le groupe considère que les utilisateurs devraient être **plus largement sensibilisés** à l'existence et au fonctionnement des API et **plus précisément informés, de manière simple et adaptée**, des conditions d'utilisation et catégories d'utilisateurs des API des réseaux sociaux qu'ils utilisent, de manière à permettre une exploitation **transparente**, respectueuse des droits des utilisateurs des réseaux et sécurisée pour les exploitants de données.

Le groupe juge également qu'**un accès simplifié et gratuit** aux bases de données comprenant les **données rendues publiques sur les réseaux sociaux par leurs utilisateurs**, devrait être accordé **aux acteurs de la recherche publique** et aux porteurs de projets de recherche scientifique financés par les pouvoirs publics ou commandés par les pouvoirs publics dans l'exercice de leurs missions de service public.

Une proposition de directive européenne en date du 14 septembre 2016[32], prévoit d'introduire une exception dite de *data mining «pour les reproductions et extractions effectuées par des organismes de recherche, en vue de procéder à une fouille de textes et de données sur des œuvres ou autres objets protégés auxquels ils ont légitimement accès à des fins de recherche scientifique»*. La loi pour une République

31. https://twitter.com/fr/tos article 4

32. COM(2016) 593 final, 2016/0280(COD) du Parlement Européen et du Conseil sur le droit d'auteur dans le marché unique numérique

Numérique[33] a introduit une exception aux droits des créateurs et producteurs de bases de données fondée sur le *data mining*[34] à des fins de recherche, mais en a limité le champ d'application aux fouilles des données *« incluses ou associées aux écrits scientifiques »*.

Le groupe considère que pour permettre un accès simplifié et gratuit aux données voulues publiques par les utilisateurs des réseaux sociaux, l'exception dite de *data mining*[35] pourrait être utilement élargie, et sa portée au moins alignée sur le projet de texte européen.

Les possibilités d'accès aux données étant circonscrites, les usages des données seront exposés dans la section suivante.

33. LOI n° 2016-1321 du 7 octobre 2016 pour une République numérique

34. Ou fouille de données

35. Article L.342-3 du code de la propriété intellectuelle

PARTIE III.

USAGES DES DONNÉES DES RÉSEAUX SOCIAUX EN SANTÉ

USAGES DES DONNÉES DES RÉSEAUX SOCIAUX EN SANTÉ

Le groupe de travail a distingué deux grands usages des réseaux sociaux en santé : la communication ciblée et la recherche.

A. LA COMMUNICATION CIBLÉE

La communication ciblée, très présente sur les réseaux sociaux généralistes, prend deux formes : les contenus viraux et les campagnes publicitaires.

Dans les deux cas, un acteur désireux de faire passer un message auprès d'une population cible, utilise un algorithme qui évalue à qui et comment adresser les messages.

La différence entre ces deux types de communication réside dans le fait que le contenu viral est spontanément relayé par l'adhésion des utilisateurs, tandis que les campagnes publicitaires sont mises en avant par le poids du budget affecté.

Nous avons répertorié trois principaux cas d'usage de communication ciblée en santé.

1. LES CAMPAGNES D'INFORMATION DE SANTÉ PUBLIQUE

Les campagnes d'information des autorités de santé utilisent aujourd'hui très régulièrement les réseaux sociaux comme un canal de diffusion complémentaire des médias traditionnels. Des community managers, aujourd'hui intégrés aux équipes de communication adaptent les formats et les stratégies pour rendre les campagnes les plus efficaces possibles.

2. LA PROPOSITION DE BIENS ET SERVICES DE SANTÉ À DESTINATION DES PATIENTS

La communication ciblée est aujourd'hui utilisée pour véhiculer une publicité sponsorisée via les grandes plateformes (Facebook/Twitter) souvent à destination de patients atteints de pathologies chroniques.

C'est notamment le cas des services numériques comme les objets connectés qui se retrouvent parfois quotidiennement dans le fil d'actualité des patients français aujourd'hui.

3. LE RECRUTEMENT DE PARTICIPANTS POUR UN MOUVEMENT/UNE ÉTUDE/ UN PROGRAMME DE SANTÉ

Un lien diffusé au bon moment et aux bonnes personnes suffit pour constituer une équipe de volontaires prêts à participer à une étude. La communauté peut être créée pour une étude, à grande échelle, ou pour une série d'études.

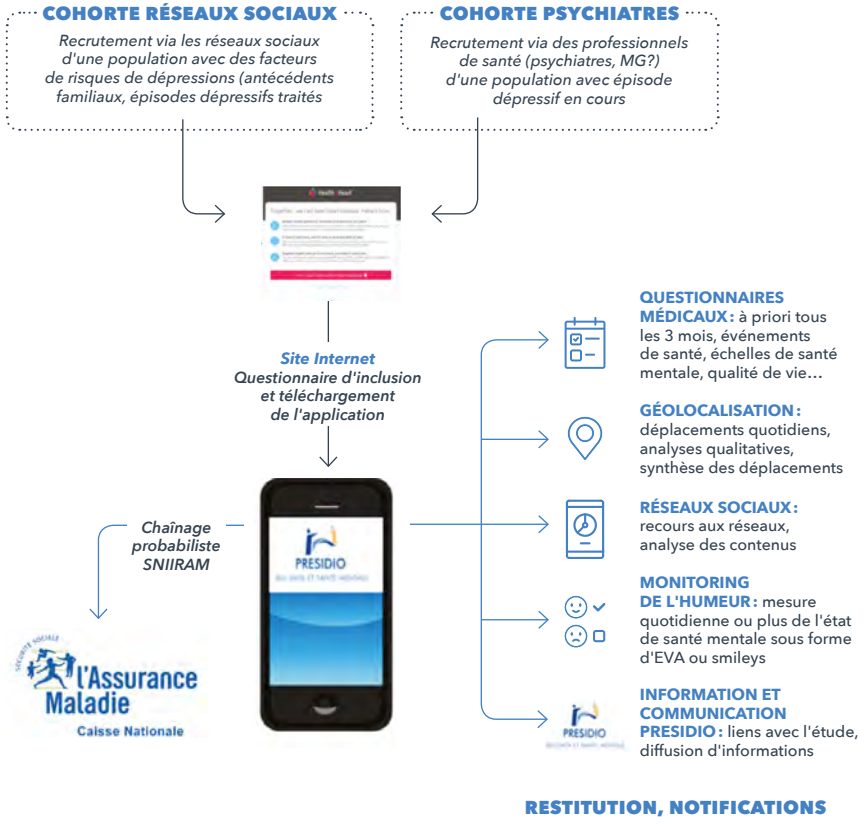
Des volontaires sont ainsi recrutés notamment pour :

- la réalisation d'études **au niveau mondial** (par ex. l'étude «World Diabetes Distress study» menée par l'INSERM) ;
- la création de réseaux de patients prêts à participer à des études scientifiques (par ex. MoiPatient, Carenity) ; et
- l'évaluation de la qualité des services de soins (par ex. CareAdvisor).

Nous pouvons ici citer le projet PRESIDIO[36] qui a pour ambition de créer, via les réseaux sociaux, une cohorte de plusieurs milliers de patients présentant des facteurs de risques de dépression ou des symptômes dépressifs. Les patients de cette «e-cohorte» seront suivis pendant plusieurs mois et les données des patients inclus seront collectées et organisées dans un cloud sécurisé dédié au projet.

36. https://presidio.fr/

Présentation du projet Presidio

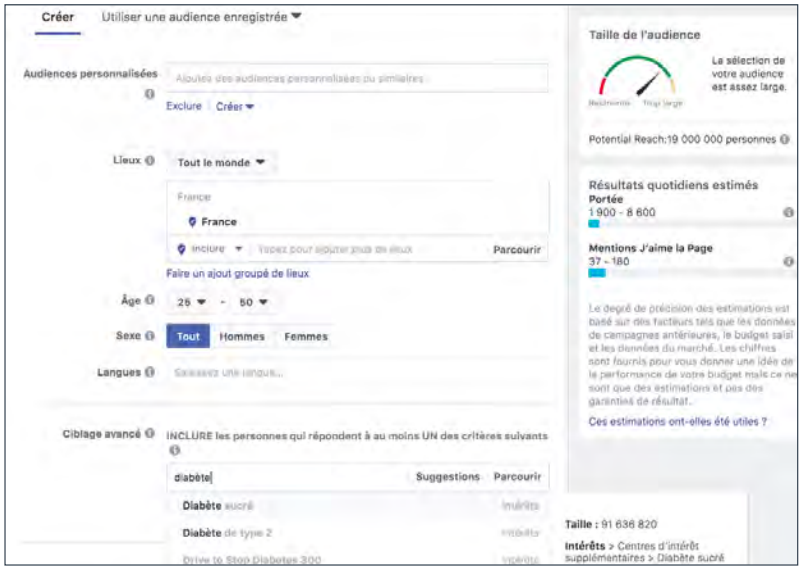


4. OBSERVATIONS DU GROUPE DE TRAVAIL

Le groupe a noté que les facultés de ciblage des utilisateurs proposées par les réseaux sociaux étaient les mêmes, quels que soient le demandeur et la finalité de sa demande (publicitaire ou équipe de recherche).

Ce système générique pour l'heure ne permet pas de cibler de manière précise une population, notamment en fonction d'une pathologie, de traitements administrés et/ou d'antécédents médicaux particuliers. Si la plateforme Facebook permet de cibler une population selon les centres d'intérêts des utilisateurs parmi lesquels peuvent figurer une ou plusieurs pathologies, ce système s'avère imparfait puisque certaines catégories de « centre d'intérêts », comme les pathologies précisément, apparaissent et disparaissent au fil du temps. (Voir Figure ci-dessous)

Ciblage à des fins de recrutement de patients



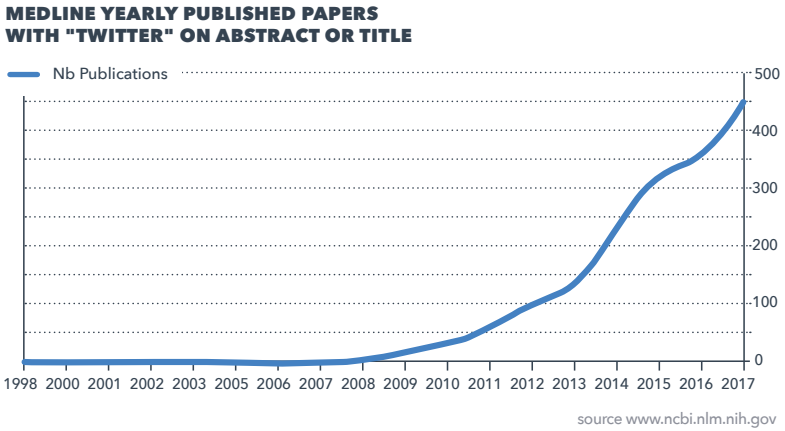
Au vu de ces constatations, le groupe considère qu'il n'est pas souhaitable que les réseaux sociaux puissent offrir à tout utilisateur de leurs services professionnels des outils de ciblage avancé, selon les pathologies ou traitements suivis par les utilisateurs, par exemple. Il estime néanmoins que le développement de **partenariats entre les pouvoirs publics et les réseaux sociaux dans le cadre desquels des modules de ciblage avancé seraient mis à la disposition des pouvoirs publics**, pourraient permettre une **diffusion efficace des campagnes publiques de prévention ou le recrutement de volontaires pour des projets de recherche publique** ou des projets conduits pour le compte des pouvoirs publics dans l'exercice de leurs missions de service public.

B. OBSERVATION DE POPULATIONS / RECHERCHE

1. LES PROJETS DE RECHERCHE CONDUITS SUR LES DONNÉES DE RÉSEAUX

Lors des différents échanges et entretiens conduits par le groupe, nous avons souvent entendu que les données de réseaux sociaux seraient difficilement exploitables puisqu'incomplètement qualifiées ou contextualisées (Est-ce bien le patient qui s'exprime ? Le contexte dans lequel est révélée l'information est-il bien identifié ? Les données relatives au patient (par ex. âge, sexe) ou à l'évènement (délai de survenue des effets indésirables, durée de l'effet déclaré, effet de l'arrêt d'un traitement, etc.) sont inconnues ou incertaines.

Ces limites certaines à l'exploitation exclusive des données de réseaux sociaux pour en déduire des résultats à l'échelle individuelle d'un patient ne font cependant pas obstacle à l'exploitation de ces données à l'échelle populationnelle. Le groupe a ainsi pu constater que la multitude des communications des patients décrivant sur les réseaux leur expérience de la maladie et des traitements était de plus en plus exploitée pour le suivi en temps réel de problématiques sanitaires.



source www.ncbi.nlm.nih.gov

Dans ce cadre, de nombreuses équipes de recherche à travers le monde se sont attelées à développer des modèles d'analyse pour apporter des réponses aux problématiques suivantes :

• **Le suivi des épidémies :** Ce suivi est notamment permis par les tweets géolocalisables. Les chercheurs de l'université Northeastern à Boston sont ainsi parvenus à développer un algorithme de prédiction de la propagation de la grippe via les tweets. Les résultats de leurs algorithmes concordent dans 70% à 90% des cas avec les données récoltées par les *Centers for Disease Control and Prevention*[37].

• **L'identification des effets indésirables des médicaments :** Cette problématique représente un enjeu majeur de santé publique. En France, uniquement 5% à 6% des effets indésirables (EI) sont déclarés aux autorités de santé. Ce manque d'information a poussé un nombre important d'équipes de recherche à mettre en place des solutions d'analyses de Big Data sur les réseaux sociaux afin d'identifier automatiquement les déclarations d'EI. C'est le cas notamment du projet Vigi4MED, financé par l'Agence Nationale de Sécurité du Médicament (ANSM), qui a mis en place un algorithme permettant d'identifier des milliers de messages postés portant sur cette thématique. Si ces études ne permettent pas de déterminer l'imputabilité d'un effet secondaire à un médicament, elles permettent en revanche d'identifier des corrélations ou des signaux traçables dans le temps.

• **L'identification des problématiques autour des parcours de soins :** Face aux différents écarts de parcours de soins, dus à un manque d'information ou de coordination, la reconstitution du parcours patient réel est devenu un enjeu médico-économique important. C'est dans ce cadre que la start-up Kap Code, par exemple, travaille sur le sujet pour les patients atteints du « syndrome de l'œil sec ».

L'objectif du projet est de donner une information quantifiée, directement issue des réseaux sociaux, et de reconstruire grâce à ces données la chronologie du parcours patient en vie réelle. Concrètement, dans le cas de ces patients, il est possible d'identifier l'ordre et les délais moyens des visites médicales par spécialité et par type de structures (hôpitaux/cliniques/cabinets libéraux etc.). Cela permet d'identifier les parcours types les plus fréquemment rencontrés. Ce cas d'usage reste toutefois

37. http://www.slate.fr/story/145746/twitter-maladie-grippe

très dépendant de ce que déclarent les patients au cours de leur prise en charge et l'accès aux données demeure complexe, ce qui ne permet pas d'envisager aujourd'hui une approche générique, applicable à l'ensemble des pathologies.

• **L'analyse de l'opinion publique sur un sujet d'actualité en santé :** Cette analyse est aujourd'hui au cœur des préoccupations des autorités sanitaires. Les réseaux sociaux deviennent un vecteur d'information non contrôlé permettant de donner la parole, sans hiérarchie des sources, à l'ensemble des Français connectés. Cette démocratisation de l'information, susceptible de générer ou d'amplifier des phénomènes de « fake news » peut avoir des conséquences délétères en santé. Le phénomène a notamment été observé avec l'extension de l'obligation de vaccination à de nouveaux produits. Les techniques actuelles de traitement automatique du langage offrent la possibilité de récolter les verbatim patients, des arguments ou des thèmes et de pouvoir les analyser pour mieux réagir face aux débats publics toujours plus dynamiques et vulnérables à la désinformation.

• **Mesurer le fardeau psychologique des maladies :** En analysant les commentaires exprimés par les patients sur les réseaux sociaux, il est possible de caractériser la charge psychologique des auteurs de commentaires. Cette identification peut aller de la classification de sentiments 100% automatique, selon des catégories prédéfinies (positif, négatif, joie, peur, rage, tristesse, indifférence etc.), ou semi-automatique à l'aide de techniques exploratoires comme le LDA (Latent Dirichlet allocation) qui peuvent être « guidée » par un utilisateur pour trouver des sujets d'intérêt.

C'est le cas de l'étude française « World Diabetes Distress study » conduite par l'unité Inserm, Paris-Sud/Paris-Saclay en épidémiologie et études populationnelles et Guy Fagherazzi, concernant les patients atteints de diabète.

Ce projet a pour but d'identifier des profils de détresse liée au diabète partout dans le monde à partir de données issues du réseau social Twitter. En effet, aujourd'hui, on en sait peu sur les profils des patients diabétiques « en vie réelle », en particulier en ce qui concerne leur profil psychologique, leurs émotions, stress, anxiété et comment ces émotions impactent leur quotidien et leur santé au long cours. Pour combler cette lacune, les chercheurs tentent d'identifier des profils de détresse liée au diabète et d'étudier leurs associations avec l'état de santé des patients, à partir de la combinaison de données numériques (« digitosome ») complexes issues des réseaux sociaux, de données cliniques/épidémiologiques récoltées par un chatbot et de données de capteurs ou objets connectés. Ces données massives sont analysées à l'aide de méthodes d'Intelligence Artificielle. Cette étude répond à deux objectifs :

- **Objectif 1.** Identification de profils de détresse liée au diabète, à partir de données textuelles issues du réseau social Twitter, partout dans le monde. Etude des déterminants socioéconomiques et environnementaux des profils de détresse, à partir des données de géolocalisation des Tweets.

- **Objectif 2.** Étude prospective des associations entre facteurs psychologiques/détresse liés au diabète, état de santé et qualité de vie des patients diabétiques partout dans le monde. Collecte d'informations cliniques/épidémiologiques, facteurs psychosociaux et échelles de qualité de vie avec « Diabot », un chatbot de suivi de patients diabétiques sur le réseau social Twitter. Suivi régulier sur une période d'au moins 3 ans.

Ce projet s'inscrit dans une philosophie d'Open Data et Open Source, ce qui signifie que les données générées et les algorithmes seront rendus accessibles. Il est à la pointe de la technologie et de la méthodologie d'analyse de données massives (Big Data) et une étude pilote réalisée en 2017 a démontré sa faisabilité. L'approche mixte, mélangeant données textuelles, cliniques et épidémiologiques va permettre d'identifier des marqueurs innovants dans le suivi du diabète.

L'état de santé d'un patient était jusqu'à présent caractérisé par quelques mesures récentes de biomarqueurs d'intérêt (glycémie, HbA1c etc.). Pour chaque patient, de milliers de données diverses sur de nombreux marqueurs d'intérêt seront à l'avenir simultanément disponibles. Ce changement est susceptible de modifier profondément la manière dont sera appréhendé le suivi d'un patient diabétique. Dans ce contexte évolutif, le projet va probablement permettre de modéliser une méthodologie d'analyse de ces données multi-sources.

2. MÉTHODOLOGIE D'ANALYSE DES DONNÉES DES RÉSEAUX

L'analyse des données des réseaux sociaux (figure 1.) ne repose pas sur un modèle capable de comprendre le langage naturel. Elle a été rendue possible grâce au développement d'outils permettant d'extraire des informations précises à partir du texte rédigé par les utilisateurs. Ces in-

Figure 1. Schéma d'analyse des données des réseaux

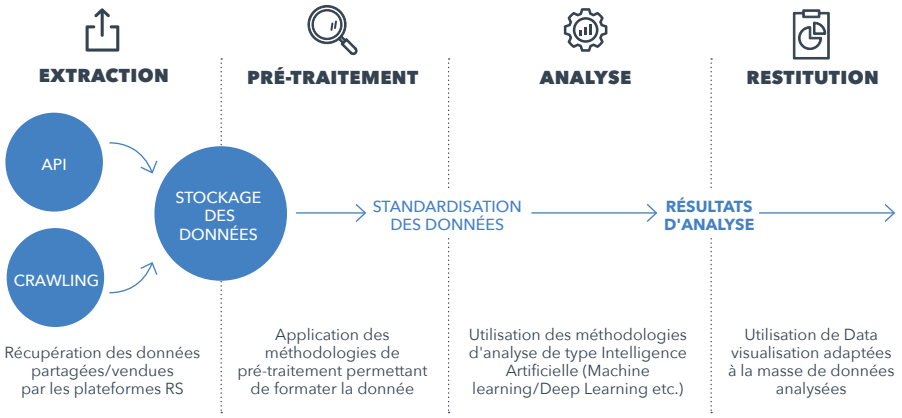
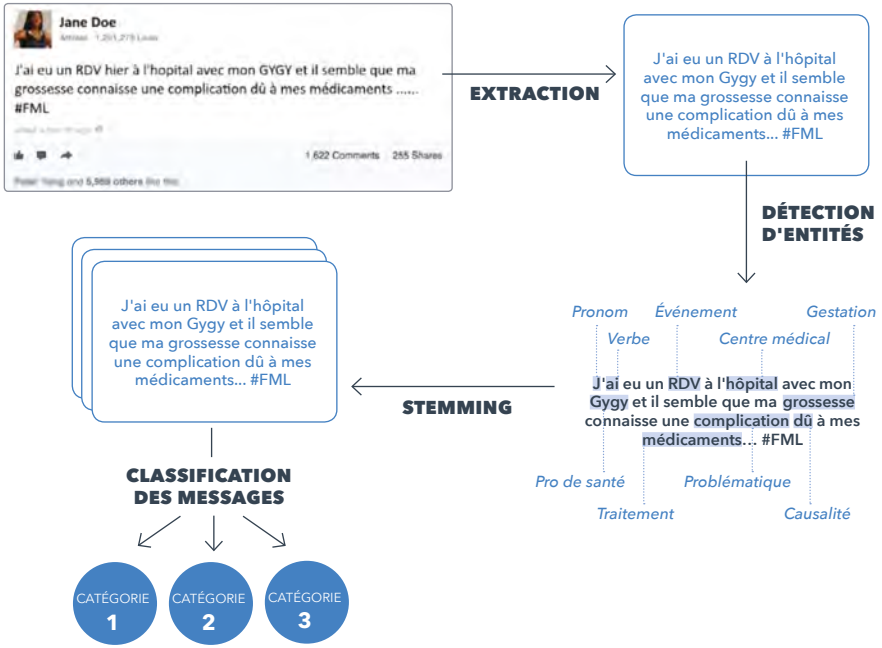


Figure 2. Les traitements subis par les messages issus des réseaux sociaux.



formations concernent l'identification des étapes du parcours patient, la classification des opinions et l'évaluation des sentiments.

Les outils d'analyse sont décrits dans le tableau figurant en Annexe I. Nous illustrons ci-contre (figure 2.) les différents traitements que peuvent subir les messages issus des réseaux sociaux.

C. PERSPECTIVES

À l'issue de cette étude, le groupe a acquis trois convictions :

- les réseaux sociaux sont devenus une source complémentaire de données de vie réelle qui doit être prise en considération, notamment par les pouvoirs publics dans le cadre de leurs missions de veille sanitaire et de prévention, comme le font déjà des autorités étrangères[38];
- ces données, précieuses pour les études populationnelles, ne permettent toutefois pas à elles seules de contextualiser un évènement à l'échelle individuelle. Pour ces usages, une approche mixte et des appariements des données des réseaux avec des données médicalisées doivent être envisagés ;
- la multiplication des études des données textuelles issues des réseaux et des données textuelles en général influera sur les pratiques de recherche qui évolueront vers une plus grande prise en compte des données de vie réelle ;
- ces évolutions ne se feront pas sans :

- une explication du raisonnement suivi et une évaluation de la performance des algorithmes (précision, compréhension, reproductibilité, etc.).

- une réflexion approfondie sur la sécurité des données partagées sur les réseaux. Cette réflexion doit être celle des réseaux eux-mêmes, des pouvoirs publics mais aussi des personnes concernées. Une juste information sur les données «détenues» par les réseaux sociaux et la sécurité qui peut raisonnablement être attendue, doit permettre à chacun de prendre une décision éclairée sur les espaces au sein desquels il souhaite effectivement partager ses données.

38. La FDA (Food and Drug Administration) ou la Société américaine de cardiologie par exemple

4 PROPOSITIONS pour une exploitation responsable des données de santé générées par les patients sur les réseaux sociaux dans l'intérêt des patients et le respect de leurs droits.

1. Favoriser l'exercice effectif du droit à la **portabilité** des données de santé générées sur les réseaux sociaux pour permettre l'exploitation de **toutes** les données partagées, **à l'initiative et sous le contrôle des utilisateurs.**

• Les utilisateurs devraient pouvoir «récupérer» leurs données organisées selon les thématiques qui ont généré leur activité (dont la santé) et non selon les catégories d'activités propres aux réseaux sociaux (publications, commentaires, mentions «J'aime» etc.) et les données de l'activité d'un compte devraient pouvoir être partiellement portables, thématique par thématique.

• Des fonctionnalités devraient être mises à la disposition des utilisateurs pour leur permettre de transférer directement leurs données à un porteur de projet de recherche via, par exemple, des outils simples comme des appels au partage de données, similaires aux outils existants d'appel aux dons[39].

Le droit à la portabilité des données est prévu par l'article 20 du RGPD. Il permet aux personnes concernées – ici les utilisateurs – de recevoir du responsable du traitement les données qui les concernent et qu'elles lui ont fournies. Cette transmission doit se faire dans un format qui permet la réutilisation de ces informations. L'article 20 prévoit également au bénéfice des personnes concernées, la faculté de transmettre leurs données à un tiers, responsable d'un second traitement de données. Elles peuvent même demander que ces données soient directement transmises par le premier responsable de traitement, au second.

2. **Sensibiliser** les citoyens à l'existence d'API et de leur fonctionnement. Les **informer, précisément et de manière simple et adaptée,** des conditions d'utilisation et catégories d'utilisateurs des API des réseaux sociaux qu'ils utilisent, pour permettre une exploitation **transparente**, respectueuse des droits des utilisateurs des réseaux et **sécurisée** pour les exploitants de données.

«L'API peut être résumée à une solution informatique qui permet à des applications de communiquer entre elles et de s'échanger mutuellement des services ou des données. Il s'agit en réalité d'un ensemble de fonctions qui facilitent, via un langage de programmation, l'accès aux services d'une application[40]».

3. Permettre un **accès simplifié et gratuit** aux bases de données, comprenant les **données rendues publiques sur les réseaux sociaux par leurs utilisateurs,** pour les **acteurs de la recherche publique** et les porteurs de projets de recherche scientifique financés par les pouvoirs publics ou commandés par les pouvoirs publics dans l'exercice de leurs missions de service public. À cette fin, l'exception légale dite de *data mining*[41] pourrait être élargie.

L'exception actuelle de data mining limite les droits d'exploitation qui permettent aux producteurs et créateurs de bases de données de contrôler et conditionner les accès aux bases qu'ils ont produites ou créées. Elle est toutefois limitée aux fouilles de données «incluses ou associées aux écrits scientifiques[42]».

4. Développer des **partenariats entre les pouvoirs publics et les réseaux sociaux** et, dans ce cadre, **mettre à la disposition des pouvoirs publics des modules de ciblage avancé** pour permettre une **diffusion efficace des campagnes publiques de prévention** ou **le recrutement de volontaires pour des projets de recherche publique** ou des projets conduits pour le compte des pouvoirs publics dans l'exercice de leurs missions de service public.

Le groupe a noté que les facultés de ciblage des utilisateurs proposées par les réseaux sociaux étaient les mêmes quels que soient le demandeur et la finalité de sa demande (publicitaire ou équipe de recherche). Ce système générique - pour l'heure - ne permet pas de cibler de manière précise une population, notamment en fonction d'une pathologie, de traitements donnés et/ou d'antécédents médicaux particuliers

39. https://fr-fr.facebook.com/help/990087377765844?helpref=faq_content

40. https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203559-api-application-programming-interface-definition-traduction/

41. Article L.342-3 du code de la propriété intellectuelle

42. Article L.342-3 du code de la propriété intellectuelle

ANNEXE I : LES OUTILS D'ANALYSE DES DONNÉES DES RÉSEAUX

TYPE	FONCTION	USAGE	PRINCIPE
PRÉPARATION	**Lemmatisation**	Il s'agit de **remplacer chaque mot d'un texte par sa racine**, par ex. la phrase « je suis fatiguée des médicaments trop chers » devient « je être fatigué des médicament trop cher ». Le but de cette transformation est de **rendre identique les mots indépendamment de leurs dérivations.**	Utilisation de dictionnaires ou règles de suffixes-préfixes communes
	Classification grammaticale (part of speech tagging)	Identification des phrases et **catégories de mots** (nom, pronom, verbe, adjectif, etc.) pour pouvoir identifier les relations entre les mots ou cibler une catégorie particulière	Décomposition des textes en phrases en utilisant les symboles de ponctuation en fin de phrase et la fréquence de transitions des mots via de **modèles de markov cachés**[43]
	Réduction du bruit	**Suppression des mots trop fréquents**	Élimination **des mots vides** à partir d'un **dictionnaire de « stop words »**[44]
	Correction orthographique	Identifier des **mots mal écrits** et les remplacer pour leur **version correcte**	Utilisation des indexes (par ex. Apache Lucene) capables de trouver les mots les plus proches **en distance levenshtein**[45] dans un dictionnaire par rapport à un mot qui ne l'est pas
IDENTI-FICATION DES ATTRIBUTS (feature engineering)	**Détection d'entités**	Identification des mentions des « entités données » par un référentiel ou **ontologie**[46], par ex. médicaments, lieux, symptômes	Comparaison avec des dictionnaires tels que l'**UMLS**[47]
	« Sacs des mots »	**Transformation du texte** à analyser en un vecteur de grande dimensionnalité > 1K qui peut être utilisé par des **algorithmes d'apprentissage supervisé ou non supervisé**	Remplacement de chaque texte à analyser par un **vecteur** ayant **une dimension pour chaque mot dans le corpus**. La valeur de chaque mot est **le nombre d'apparitions du mot dans le texte** en question.

43. *« Un modèle de Markov caché (MMC, terme et définition normalisés par l'ISO/CÉI [ISO/IEC 2382-29:1999]) – en anglais : hidden Markov model (HMM) –, ou plus correctement (mais non employé) automate de Markov à états cachés, est un modèle statistique dans lequel le système modélisé est supposé être un processus markovien de paramètres inconnus. Contrairement à une chaîne de Markov classique, où les transitions prises sont inconnues de l'utilisateur mais où les états d'une exécution sont connus, dans un modèle de Markov caché, les états d'une exécution sont inconnus de l'utilisateur (seuls certains paramètres, comme la température, etc. sont connus de l'utilisateur). Les modèles de Markov cachés sont massivement utilisés notamment en reconnaissance de formes, en intelligence artificielle ou encore en traitement automatique du langage naturel. »*
https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_Markov_cach%C3%A9

44. *« En recherche d'information, un mot vide (ou stop word, en anglais) est un mot qui est tellement commun qu'il est inutile de l'indexer ou de l'utiliser dans une recherche. En français, des mots vides évidents pourraient être "le", "la", "de", "du", "ce", … »*
https://fr.wikipedia.org/wiki/Mot_vide

45. *« La distance de Levenshtein est une distance, au sens mathématique du terme, donnant une mesure de la différence entre deux chaînes de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaîne à l'autre.Elle a été proposée par Vladimir Levenshtein en 1965. Elle est également connue sous les noms de distance d'édition ou de déformation dynamique temporelle, notamment en reconnaissance de formes et particulièrement en reconnaissance vocale.*
Cette distance est d'autant plus grande que le nombre de différences entre les deux chaînes est grand. La distance de Levenshtein peut être considérée comme une généralisation de la distance de Hamming. On peut montrer en particulier que la distance de Hamming est un majorant de la distance de Levenshtein. »
https://fr.wikipedia.org/wiki/Distance_de_Levenshtein

46. *« En informatique et en science de l'information, une ontologie est l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné. Plus simplement, on peut aussi dire que l'"ontologie est aux données ce que la grammaire est au langage".*
Le terme est utilisé par analogie avec le concept philosophique, d'ontologie (de onto-, tiré du grec ὤν, ὄντος "étant", participe présent du verbe εἰμί "être") qui est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe.
Les concepts sont organisés dans un graphe dont les relations peuvent être :
• des relations sémantiques ;
• des relations de subsomption.
L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné, qui peut être réel ou imaginaire.
Les ontologies sont employées dans l'intelligence artificielle, le Web sémantique, le génie logiciel, l'informatique biomédicale ou encore l'architecture de l'information comme une forme de représentation de la connaissance au sujet d'un monde ou d'une certaine partie de ce monde. Les ontologies décrivent généralement :
• les individus : les objets de base ;
• les classes : ensembles, collections, ou types d'objets1 ;
• les attributs : propriétés, fonctionnalités, caractéristiques ou paramètres que les objets peuvent posséder et partager ;
• les relations : les liens que les objets peuvent avoir entre eux ;
• les événements : changements subis par des attributs ou des relations ;
• une métaclasse (web sémantique) : des collections de classes qui partagent certaines caractéristiques. »
https://fr.wikipedia.org/wiki/Ontologie_(informatique)

47. Unified Medical Language System, voir notamment https://www.nlm.nih.gov/research/umls/quickstart.html

IDENTIFICATION DES ATTRIBUTS (feature engineering)	**Word2Vec**[48]	Transformation du texte à analyser en un vecteur avec environ 300 dimensions. Ces vecteurs sont calculés en amont par rapport à un grand corpus de textes et capturent la sémantique de chaque mot. **Le principe est que deux mots sont interchangeables, par ex. joli, beau ont des vecteurs similaires.** Ces vecteurs peuvent aussi être composés par des opérations arithmétiques, par ex. Vecteur «Roi» - Vecteur «Homme» + Vecteur «Femme» - Vecteur «Reine»	Un réseau de neurones avec une couche interne est entraîné sur un grand corpus à identifier le mot qui a été retiré aux phrases (CBOW). A la fin de l'entraînement on utilise les poids des neurones internes pour construire un vecteur pour chaque mot.
APPRENTISSAGE SUPERVISÉ	**Identification des sentiments**	Classifier des phrases en fonction **des sentiments binaire** (positif-négatif) ou **multiples** (colère, joie, tristesse, etc.)	Construction d'un jeu de phrases classifiées manuellement par une ou plusieurs personnes qui est ensuite séparé en entraînement et test. Un modèle, par ex. SVM[49] ou Random Forest[50], se base sur la **vectorisation des textes** et **apprend par le jeu d'entraînement**. Ses performances sont ensuite testées avec le jeu de test.
	Identification des thèmes	Identifier si **une phrase évoque un sujet ou une opinion particulière** (par ex. anti vaccination)	Même méthodologie que pour l'identification des sentiments.
	Création de signaux	Création des **signaux** ou **(time series)** qui peuvent être utilisés pour **la détection des événements ou alertes** : par ex. effets indésirables des médicaments.	Utilisation des techniques d'**identification des thèmes** et/ou d'**identification des entités** et calcul du **nombre d'occurrences** dans le temps (il faut que les textes soient datés). Ce traitement se poursuit par des techniques d'analyse des signaux comme les **modèles de Markov cachés**
APPRENTISSAGE NON SUPERVISÉ	**Clustering**	Regroupement des **textes avec une sémantique proche** pour faciliter leur classification par une personne.	Vectorisation des textes (sacs de mots ou word2 Vec[51]) et classification avec un modèle de clustering, par ex. K-means[52]
	Identification de topiques	Analyse des discours pour identifier **les mots qui forment ensemble les sujets abordés**. Le résultat est une identification des sujets, des mots principaux des sujets et de la présence des sujets dans les textes.	**Algorithme LDA**[53] (Latent Dirichlet allocation)

48. *«En intelligence artificielle et en apprentissage machine, Word2vec est un groupe de modèles utilisé pour le plongement lexical (word embedding). Ces modèles ont été développés par une équipe de recherche chez Google sous la direction de Tomas Mikolov (en). Ce sont des réseaux de neurones artificiels à deux couches entraînés pour reconstruire le contexte linguistique des mots. La méthode est implémentée dans la bibliothèque Python Gensim1.»*
https://fr.wikipedia.org/wiki/Word2vec

49. *«Les machines à vecteurs de support ou séparateurs à vaste marge (en anglais support vector machine, SVM) sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discriminationnote 1 et de régression. Les SVM sont une généralisation des classifieurs linéaires.*
Les séparateurs à vaste marge ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la théorie de Vapnik-Chervonenkis. Ils ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyperparamètres, leurs garanties théoriques, et leurs bons résultats en pratique.
Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur, finance1…). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un modèle de mélanges gaussiens.»
https://fr.wikipedia.org/wiki/machine_%c3%a0_vecteurs_de_support

50. *«Les forêts d'arbres décisionnels1 (ou forêts aléatoires de l'anglais random forest classifier) ont été formellement proposées en 2001 par Leo Breiman et Adèle Cutler.*
Elles font partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de bagging. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.»
https://fr.wikipedia.org/wiki/For%C3%AAt_d%27arbres_d%C3%A9cisionnels

51. Voir note n°6

52. *«Le partitionnement en k-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances.*
Il existe une heuristique classique pour ce problème, souvent appelée méthodes des k-moyennes, utilisée pour la plupart des applications. Le problème est aussi étudié comme un problème d'optimisation classique, avec par exemple des algorithmes d'approximation. Les k-moyennes sont notamment utilisées en apprentissage non supervisé où l'on divise des observations en k partitions. Les nuées dynamiques sont une généralisation de ce principe, pour laquelle chaque partition est représentée par un noyau pouvant être plus complexe qu'une moyenne. Un algorithme classique de k-means est le même que l'algorithme de quantification de Lloyd-Max.»
https://fr.wikipedia.org/wiki/K-moyennes

52. *«Dans le domaine du traitement automatique des langues, l'allocation de Dirichlet latente (de l'anglais Latent Dirichlet Allocation) ou LDA est un modèle génératif probabiliste permettant d'expliquer des ensembles d'observations, par le moyen de groupes non observés, eux-mêmes définis par des similarités de données.»*
https://fr.wikipedia.org/wiki/Allocation_de_Dirichlet_latente

ANNEXE II : PERSONNES ENTENDUES ET REMERCIEMENTS

Au cours de ses travaux, le groupe de travail s'est entretenu avec :

- **Madame Michèle Arnoé,**
  Directrice Innovation & Croissance, société IQVIA
- **Madame Anne Buisson,**
  Directrice adjointe de l'Association François Aupetit (AFA) Crohn RCH France
- **Madame Yvanie Caillé,**
  Directrice de l'Institut National des Données de Santé, fondatrice de l'Association Renaloo
- **Monsieur Guy Fagherazzi, PhD,**
  Chercheur en épidémiologie (Digital and Diabetes Epidemiology), CESP (Centre de Recherche en Epidémiologie et Santé des Populations) UMR 1018, Inserm, Université Paris Sud-Paris Saclay, Gustave Roussy
- **Monsieur Gilles Garnie,**
  Vice-président de l'association France Psoriasis
- **Madame Marjolaine Hering,**
  Association Française de l'Eczéma
- **Monsieur Guillaume Jeannerod,**
  Directeur Général de la société Epiconcept
- **Monsieur Pascal Nieters,**
  Post-doctorant en neuro informatique Institut für Kognitionswissenschaft (Institute of Cognitive Science), Osnabrück, Allemagne
- **Madame Dominique Noël,**
  Présidente de Festival de la communcation santé
- **Madame Lise Radoszycki,**
  Directrice, Data Science, société Carenity
- **Monsieur David Réguer,**
  Directeur Général de la société RCA Factory
- **Madame Laura Sablone,**
  Chef de projet de recherche, Association Seintinelles
- **Monsieur Eric Salat,**
  patient expert - Codirecteur diplôme Université des patients
- **Monsieur Claude Touche,**
  Président de la société eVeDrug
- **Monsieur Alain Veuillet,**
  Directeur des Relations avec les Pharmaciens & du Développement Officinal - Groupe Pierre Fabre
- **Monsieur Pascal Vilain,**
  Épidémiologiste à Santé publique France (Cire Océan Indien).

Le groupe de travail tient à les remercier du temps qu'ils ont consacré au projet et de la générosité avec laquelle ils ont fait part de leurs expérience et connaissances.

Il tient également à remercier Monsieur Christophe Guillot, Directeur digital Solutions Partners - Groupe Pierre Fabre qui a initié cette réflexion sur les patients et leurs données de santé.

QUELQUES REPÈRES BIBLIOGRAPHIQUES

PUBLICATIONS SCIENTIFIQUES

- Valérie-Laure Benamou, L'extension du domaine de la donnée, in Legipresse n°359, Avril 2018, Chron. & Opinions p 197
- Blei, David M.; Ng, Andrew Y.; Jordan, Michael I (January 2003). Lafferty, John, ed. Latent Dirichlet Allocation, Journal of Machine Learning Research. 3 (4–5): pp. 993–1022.
- Jeremy A. Greene, Niteesh K. Choudhry, Elaine Kilabuk, William H. Shrank, Online Social Networkin by Patients with Diabetes: A Qualitative Evaluation of Communication with Facebook, Journal of General Internal Medecine, March 2011, Vol. 26, Issue 3, pp 287-292
- Zellig S. Harris (1954) Distributional Structure, Word, 10:2-3, 146-162, DOI: 10.1080/00437956.1954.11659520
- Christophe F. Herbert, Alin Brunet, Psychiatrie et Facebook : Illustration de l'utilisation des sites sociaux au lendemain d'un trauma, L'information psychiatrique, 2010/9 Volume 86, pages 745 à 752
- Klein, A., Sarker, A., Rouhizadeh, M., O'Connor, K., & Gonzalez, G. (2017). Detecting personal medication intake in Twitter: An annotated corpus and baseline classification system. BioNLP 2017, 136-142.
- Joseph Josy Lévy, Christine Thoër, Internet et Santé : acteurs, usages et appropriations, PUQ, 1er janvier 2012
- Mao, J. J., Chung, A., Benton, A., Hill, S., Ungar, L., Leonard, C. E., ... & Holmes, J. H. (2013). Online discussion of drug side effects and discontinuation among breast cancer survivors. Pharmacoepidemiology and drug safety, 22(3), 256-262.
- Déborah Mascalzoni, Angelo Paradiso, Matts Hansson, Rare disease research: Breaking the privacy barrier, Applied and Translational Genomics 3 (2014) 23-29
- Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean (Sep 2013) Efficient Estimation of Word Representations in Vector Space arXiv:1301.3781
- Andrew G. Reece, Christopher M. Danforth, Instagram photos reveal predictive markers of depression,arXiv:1608.03282v2 [cs.SI],
- https://arxiv.org/ftp/arxiv/papers/1608/1608.03282.pdf
- Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. Drug safety, 39(3), 231-240.
- Daniel Ramage,Hidden Markov Models Fundamentals, CS229 Section Notes (Dec 2007), http://cs229.stanford.edu/section/cs229-hmm.pdf
- Quentin Roset, Réseaux sociaux de santé et enjeux pour l'industrie pharmaceutique, Thèse, 2016, n°106, Université Claude Bernard Lyon 1, Faculté de Pharmacie
- Abou Taam, M., Rossard, C., Cantaloube, L., Bouscaren, N., Roche, G., Pochard, L., ... & Bagheri, H. (2014). Analysis of patients' narratives posted on social media websites on benfluorex's (Mediator®) withdrawal in France. Journal of clinical pharmacy and therapeutics, 39(1), 53-55.
- Van de Loo, J., De Pauw, G., & Daelemans, W. (2016). Text-based age and gender prediction for online safety monitoring. Comput. Linguistics Netherlands, 5(1), 46-60.

PRESSE

- F. Boissier, Pharmacovigilance: surveiller les réseaux sociaux pourrait générer des signaux d'alerte précoces, 18 octobre 2016, https://www.ticpharma.com/story.php?cible=Innovations&story=50
- B. Broderick, Facebook on Board With FDA Opioid Meeting, CEO says, Bloomberg Law, April 12, 2018, https://www.bna.com/facebook-board-fda-n57982090974/

- F. Yoo Chee, D. Rodrigues, L'UE s'impatiente et menace Facebook et Twitter de sanctions, 20 septembre 2018, https://www.usinenouvelle.com/article/l-ue-s-impatiente-et-menace-facebook-et-twitter-de-sanctions.N744214
- J. Demey, Comment l'ordinateur peut prévenir le suicide, 3 mars 2018, https://www.lejdd.fr/societe/sciences/comment-lordinateur-peut-prevenir-le-suicide-3585359
- B. Gualliard, Epidémie on line : comment les réseaux sociaux aident à surveiller les épidémies, 13 septembre 2018, http://theconversation.com/epidemie-on-line-comment-les-reseaux-sociaux-aident-a-surveiller-la-grippe-et-les-epidemies-102227
- C. Guabert, Lévothyrox, le rapport qui pointe les failles du système dans la gestion de crise, 4 septembre 2018, https://www.sciencesetavenir.fr/sante/levothyrox-le-rapport-qui-pointe-les-failles-du-systeme-dans-la-gestion-de-la-crise_127207
- Homanides.fr, Newsroom, Comment un algorithme détecte les comportements dépressifs sur Instagram, 22 Août 2016, https://humanoides.fr/algorithme-instagram-depression-photos/
- S. Liao, Facebook could face EU sanctions if it doesn't change its terms of service, 19 septembre 2018, https://www.theverge.com/2018/9/19/17880348/facebook-european-union-fines-sanctions-terms-of-service-changes
- S. Meredith, Here's Everything You Need to Know about Cambridge Analytica Scandal 21 March 2018, https://www.cnbc.com/2018/03/21/facebook-cambridge-analytica-scandal-everything-you-need-to-know.html
- R. Moreaux, Les réseaux sociaux de plus en plus scrutés par la recherche médicale, 29 juin 2018, https://www.ticpharma.com/story.php?story=649
- J.-Y. Nau, Santé et Big Data : l'échec de Google Flu® ne doit pas cacher la forêt, 1er juillet 2016, http://www.slate.fr/story/120441/sante-big-data-echec-google-flur

ÉTUDES, LIVRES BLANCS, RAPPORTS, LIGNES DIRECTRICES

- B. Bégaud, D. Polton, F. von Lennep, Les données de vie réelle, un enjeu majeur pour la qualité des soins et la régulation du système de santé, Rapport réalisé à la demande de Madame La Ministre de la santé Marisol Touraine, mai 2017, https://solidarites-sante.gouv.fr/IMG/pdf/rapport_donnees_de_vie_reelle_medicaments_mai_2017vf.pdf
- Fing, MesInfos santé vers un Blue Button à la française, avril 2016, http://mesinfos.fing.org/wp-content/uploads/2016/05/MesInfos_Sante_Pages.pdf
- Article 29 Data Protection Working Party, Guidelines on the Right to Data Portability, adopted on 13 December 2016 As last Revised and Adopted on 5 April 2017, 16/EN, WP 242 rev.01
- Article 29 Data Protection Working Party, Guidelines on Transparency under Regulation 2016/679, adopted on 29 November 2017, As Last Revised and Adopted on 11 April 2018, 17/EN, WP260 rev. 01
- Article 29 Data Protection Working Party, Guidelines on Consent under Regulation 2016/679, Adopted on 28 November 2017 As Last Revised and Adopted on 10 April 2018

À PROPOS DU
HEALTHCARE DATA INSTITUTE

Créé en 2014, le Healthcare Data Institute est le premier Think Tank international consacré au Big Data dans le domaine de la santé, il agit comme un catalyseur d'idées et de projets autour du Big Data dans l'écosystème santé.

Le conseil d'administration du Healthcare Data Institute rassemble des représentants d'Aviesan, de McKinsey & Company, du groupe Elsan, d'IQVIA, de Sanofi, d'Orange Healthcare, de Pons & Carrère, du CRI, du CEA et d'IVBAR France.

HEALTHCARE
DATA INSTITUTE

21, rue Jasmin
75016 PARIS - FRANCE

CONTACT
Quentin ROSET
office@healthcaredatainstitute.com
+33 (0)1 42 21 19 59

@HCDATAINSTITUTE
healthcaredatainstitute.com