DOCUMENTER TWITTER
Défis et méthodes pour la constitution de corpus de tweets

Antonin Segault

Maître de conférences, Université Paris Nanterre, Laboratoire Dicen-IdF
antonin.segault@parisnanterre.fr

La plateforme de micro-blogues Twitter a pris une place centrale dans le paysage médiatique des dernières années. L'archivage des messages (ou tweets) qui y sont publiés présente par conséquent une importance particulière, pour les chercheurs mais aussi pour la société. Cependant, pour être exploitable, leur enregistrement nécessite de prendre en compte un certain nombre de spécificités de la plateforme et des contenus qui y circulent. Dans cet article, nous revenons sur les méthodes de collecte existantes, afin de déterminer les types de données qu'elles permettent effectivement de capturer et ceux qui y échappent. Nous examinons par ailleurs l'évolution temporelle des tweets après leur publication et proposons des méthodes susceptibles de l'enregistrer. À travers la définition de ce processus de collecte, la nature documentaire des tweets et les freins à leur étude sont également interrogés.

Mots-clés : twitter, collecte de données, archivage, document, médias sociaux

In the last few years, the micro-blogging platform Twitter took a central place in the media space. Archiving the messages (or tweets) that are published on this platform is therefore an important challenge for researchers but also for the society itself. However, to be effective, this process needs to take into account some of the specificities of the platform and its contents. In this article, we analyze the existing data collection methods to assess what kind of data they do capture and what kind they do not. We also investigate the temporal evolution of tweets after their publication and propose methods to record such changes. Through the definition of this data collection process, we also question the documentary nature of tweets and the obstacles that limit their study.

Keywords: twitter, data collection, archive, document, social media

Licence CC-BY-SA

Figures emblématiques du web participatif qui naît au début des années 2000, les médias sociaux peuvent être définis comme des dispositifs numériques dédiés à la création et au partage de « contenus générés par les utilisateurs » [Kaplan & Haenlein, 2010]. Leur facilité d'utilisation a favorisé leur large adoption, bien au-delà des cercles technophiles et des usages qui leur étaient prescrits. Aujourd'hui, ces plateformes occupent une place centrale dans le monde médiatique, la communication politique, les mouvements sociaux, etc. Pour cette raison, l'étude des messages qui y circulent – en temps réel mais surtout *a posteriori* – présente un grand intérêt, pour les chercheurs en sciences sociales comme pour les membres de la société eux-mêmes. La sauvegarde des contenus publiés sur les médias sociaux s'inscrit dans la lignée des multiples projets d'archivage du web mis en œuvre depuis le milieu des années 1990 [Musiani, Paloque-Bergès, Schafer & Thierry, 2019 ; Rogers, Brügger & Milligan, 2018]. Cependant, les spécificités de ces plateformes nécessitent le développement de nouveaux outils et de nouvelles méthodes adaptées.

Dans cet article, nous nous intéresserons plus spécialement à la plateforme Twitter et à la nature des contenus qui y sont publiés. Nous reviendrons tout d'abord sur les caractéristiques de cette plateforme, ses usages actuels et les projets d'archivage la concernant. Nous proposerons ensuite une typologie des différents éléments qui composent les tweets et présenterons des méthodes permettant d'en assurer un archivage aussi exhaustif que possible. Nous nous intéresserons également à la problématique de l'évolution temporelle des contenus publiés sur Twitter, et examinerons différentes approches pour capturer ces changements à court et moyen terme. Sur la base de ces réflexions, nous proposerons de reconsidérer la nature documentaire des tweets et, finalement, interrogerons le cadre légal de leur collecte.

ARCHIVER TWITTER

Twitter est une plateforme de micro-blogues créée en 2006, permettant la publication en ligne de tweets, des messages textuels d'une longueur limitée à 280 caractères. Les tweets peuvent cependant contenir des liens hypertextes, des images fixes ou animées, des émoticônes graphiques, des mentions d'autres utilisateurs (leur pseudonyme précédé du caractère @) et des mots-clefs (hashtags ou mot-dièses, précédés du caractère #) utilisés à des fins d'indexation folksonomique [Potts, Seitzinger, Jones & Harrison, 2011]. Les tweets d'un utilisateur sont affichés dans le fil d'actualité (*timeline*) de tous les utilisateurs qui se sont abonnés (*following*) à ses publications. Un utilisateur peut également republier (retweet) ou aimer (like) la publication d'un

Licence CC-BY-SA

autre afin de la rendre visible par ses propres abonnés, permettant ainsi une diffusion virale des messages.

Créé en 2006, Twitter connaît un succès considérable, avec 330 millions d'utilisateurs actifs (au cours du dernier mois) en 2017 [Molina, 2017], atteignant le rang de onzième site le plus visité au monde en 2019 [Alexa Internet, 2019]. Cette plateforme a tout d'abord été remarquée lors de situations de catastrophes, au cours desquels des citoyens l'ont utilisée pour partager rapidement des informations vitales [Vieweg, Hughes, Starbird & Palen, 2010]. Avec l'élargissement de sa base d'utilisateurs, sont apparus des discours plus critiques, notamment en raison des rumeurs et des informations inexactes qui sont propagées dans les tweets [Starbird, Maddock, Orand, Achterman & Mason, 2014 ; Vosoughi, Roy & Aral, 2018]. Cette tendance, restée forte jusqu'à ce jour, n'a pas empêché le développement de la plateforme, y compris dans la communication politique. Le style du président Trump témoigne de l'émergence d'un « âge de Twitter », où la vérité et l'étiquette sont fréquemment malmenées [Ott, 2017]. En France, cette plateforme et les discours qui s'y déploient ont également occupé une place centrale dans les controverses politiques et électorales des dernières années [Cervulle & Pailler, 2014 ; Mercier, 2015]. Dans ce contexte, et à l'approche de plusieurs échéances électorales majeures, Twitter a interdit fin 2019 le recours aux tweets sponsorisés (payants) à caractère politique [Durupt, 2019].

Au vu de la place occupée par Twitter dans la société contemporaine, l'archivage des tweets peut sembler d'une évidente nécessité pour les besoins de la recherche mais aussi dans une perspective historique et patrimoniale [Musiani & Schafer, 2016]. Néanmoins, la quantité de données que représentent les millions de tweets publiés chaque jour constitue un obstacle considérable. Ainsi, en 2010, la Library of Congress, chargée du dépôt légal aux États-Unis, a-t-elle dévoilé un projet d'archivage de tous les tweets. Elle a reçu de la part de Twitter l'intégralité des messages publiés depuis 2006, ainsi qu'un accès aux flux de messages publiés en temps réel. Néanmoins, en 2017, la Library of Congress a mis fin à cette collecte exhaustive, évoquant notamment des difficultés liées au volume des données, au profit d'une collecte sélective, guidée par les événements d'intérêt national [Bruns, 2018 ; Library of Congress, 2017]. On retrouve ces approches d'archivage ciblé dans les collectes réalisées par l'Institut national de l'audiovisuel (INA) au titre du dépôt légal du web : près de 15 000 comptes Twitter liés à l'audiovisuel français ont été sélectionnés afin que leurs publications soient archivées (INA, 2019 ; INA, s. d.).

La collecte sélective – limitée aux publications d'une courte liste d'utilisateurs, aux messages contenant quelques mots-clefs ou provenant d'une

Licence CC-BY-SA

zone géographique réduite – permet de restreindre la taille de corpus, et ainsi d'éviter les écueils relatifs à l'enregistrement, au stockage et à la manipulation de masses de données trop importantes. Cependant, certains problèmes demeurent, qui ne sont pas liés au volume des données, mais à des caractéristiques des tweets eux-mêmes : « les sources documentaires et données numériques sont marquées par l'hétérogénéité et la multiplicité des couches d'information entre interface et machine [...] Ces sources ne sont que très difficilement lisibles et compréhensibles en dehors des dispositifs qui conditionnent leur appréhension » [Paloque-Bergès, 2016]. Afin que les corpus constitués via Twitter puissent être exploités, il s'agit donc de développer des méthodes de collecte spécifiques, adaptées à la nature des tweets.

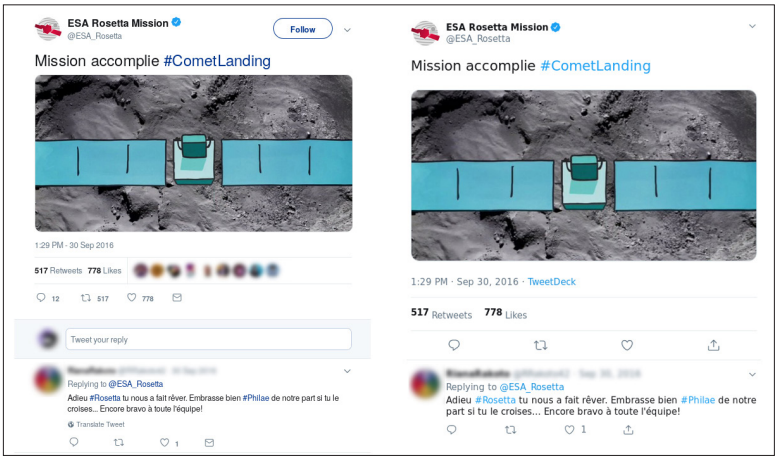
LA NATURE DES TWEETS

Un document numérique peut être défini comme la combinaison d'une structure et de données numériques : « Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son constructeur et ses lecteurs » [Pédauque, 2006, p. 45]. Appliquer ce modèle aux tweets, sans pour l'instant préjuger de leur statut de document, s'avère intéressant pour l'étude des problématiques d'archivage. En effet, la sauvegarde d'un tweet consiste alors en la collecte et la préservation des données qui le composent, mais également des structures permettant sa re-présentation ultérieure.

À première vue, un tweet est un objet à la fois simple et de taille réduite. Lorsqu'il s'affiche dans un fil d'activité, les données qui le constituent semblent se limiter à un court texte, accompagné d'un nom d'utilisateur, d'une image de profil et d'une date. Selon les cas, peuvent également apparaître des nombres de retweets et de likes, des réponses (*reply*) ainsi que des aperçus des images ou des liens attachés au tweet. Sa structure, générée par le navigateur web à partir d'instructions HTML, prend sur l'écran la forme d'un rectangle compact, où le rôle des différents éléments de contenu est rappelé par de subtiles combinaisons de format, de position et de pictogrammes. Différentes modalités d'interactions, notamment par des liens cliquables et des infobulles apparaissant au survol, enrichissent ces éléments et les relient à d'autres.

Licence CC-BY-SA

Figure 1. Captures d'écran d'un même tweet sur les sites web (à gauche) et mobiles (à droite) de Twitter : < https://twitter.com/ESA_Rosetta/status/781818209842434048 >



Pourtant, cette représentation du tweet n'en est qu'une parmi les nombreuses possibles selon les terminaux et les logiciels utilisés [Clavert, 2018]. Outre son site web *responsive* (dont l'affichage change selon les écrans), Twitter propose également un site web mobile, une application iOS (avec des interfaces différentes pour les téléphones iPhone, les tablettes iPad et les lecteurs multimédias Apple TV), deux applications pour les systèmes Android (téléphones et tablettes) et une application pour les systèmes Microsoft (téléphones, tablettes, ordinateurs, mais aussi casques de réalité virtuelle HoloLens). On trouve également de multiples applications tierces pour une large gamme de plateformes, incluant notamment les montres connectées. Ce sont donc autant de représentations graphiques différentes, n'incluant pas forcément les mêmes informations et associées à des interactions propres, qui peuvent exister pour un tweet donné. Par exemple, le site web de Twitter présente un compteur de réponses qui n'existe pas sur la version mobile, tandis que ce dernier affiche le nom de l'application qui a publié le tweet (voir figure 1). S'y ajoutent enfin des représentations informatiques, bien plus complètes, généralement structurées dans le format JSON (voir tableau 1).

Licence CC-BY-SA

Tableau 1. Extrait de la représentation informatique du tweet de la figure 1


```
{
 'in_reply_to_screen_name':None,
 'contributors':None,
 'is_quote_status':False,
 'id':781818209842434048,
 'in_reply_to_user_id_str':None,
 'retweet_count':512,
 'in_reply_to_status_id_str':None,
 'id_str':'781818209842434048',
 'coordinates':None,
 'lang':'fr',
 [...]
 'created_at':'Fri Sep 30 11:29:21 +0000 2016',
 'place':None,
 'text':'Mission accomplie #CometLanding https://t.co/82l9WBIISu'
}
```


En effet, un tweet se compose de beaucoup plus de données qu'on ne pourrait le penser au premier abord. Celles-ci relèvent de deux types : « les informations visibles à l'interface homme-machine ; les informations invisibles de programmes qui traitent les informations » [Paloque-Bergès, 2016]. Ainsi, les interfaces de programmation (API) mises à disposition des développeurs d'applications tierces permettent d'accéder, pour chaque tweet, à une trentaine d'attributs souvent « invisibles » et de complexité variée, comprenant notamment la langue du message, son origine géographique, ou encore l'application qui l'a publié [Twitter, 2019e]. S'y ajoutent également une quinzaine de métadonnées décrivant le profil de l'auteur (description, localisation, etc.). Selon les besoins, on peut également considérer qu'il faut inclure dans la sauvegarde du tweet d'autres données de contexte, relatives aux tweets précédents, aux hashtags, aux liens hypertextes ou encore aux autres comptes utilisateurs mentionnés dans le message.

En reprenant la distinction précédemment évoquée entre données et structures, on peut ainsi dresser une typologie des éléments constitutifs d'un tweet dans ces deux domaines (voir tableau 2). Pour être complète, la collecte de tweets doit permettre l'enregistrement de l'ensemble de ces éléments, et leur reproduction ultérieure de manière fidèle à l'original.

Licence CC-BY-SA

Tableau 2. Typologie des éléments constitutifs d'un tweet

Données			Structures		
Contenus affichés dans l'interface (visibles)	Métadonnées relatives au tweet (invisibles)	Données de contexte (visibles ou invisibles)	Présentation graphique (selon les terminaux)	Éléments interactifs (selon les terminaux)	Représentation informatique (JSON)
Ex. : texte du tweet, date, nom de l'auteur, etc.	Ex. : langue, application, identifiants, entités, etc.	Ex. : profils utilisateurs, précédents tweets, etc.	Ex. : apparence dans un navigateur	Ex. : liens cliquables, infobulles, vidéos, etc.	Ex. : données fournies par une API

CAPTURE DES TWEETS

Une grande variété de dispositifs de collecte des tweets a déjà été développée. Leur fonctionnement peut être rattaché à trois principales modalités de collecte de données : l'aspiration web, l'utilisation des API et la capture d'écran.

L'aspiration web consiste à télécharger une page web, ou une portion de page web, dans le but de l'archiver. Il s'agit de sauvegarder le code HTML, mais également l'ensemble des ressources nécessaires à son affichage : feuilles de styles CSS, images, scripts, polices de caractères, etc. Cette collecte peut se limiter à une seule page, mais peut également être étendue, de manière récursive, à l'ensemble des liens présents sur cette page. De nombreux aspirateurs de sites web généralistes peuvent être exploités pour la collecte de tweets, à condition de bien comprendre la structure des URL utilisées par Twitter [Blumenthal, 2019]. L'aspiration peut être effectuée à partir de pages contenant des séries de tweets (fil d'actualité, profil d'un utilisateur, page de résultats du moteur de recherche, etc.), mais aussi de la page d'un tweet seul, qui présente davantage d'informations de contexte (réponses, profils d'utilisateurs ayant retweeté ou liké, etc.). L'utilisation d'une aspiration récursive (qui suit les liens) assure l'enregistrement de toutes ces données ainsi que d'autres, telles que le profil de l'auteur et les contenus multimédias, mais élargit et complexifie fortement le corpus résultant. Par ailleurs, le résultat de la collecte peut différer légèrement de la présentation des tweets dans le navigateur, notamment parce que la majorité des aspirateurs ne peuvent simuler certaines interactions utilisateur (survol, défilement) et donc capturer leur résultat.

D'autres outils s'appuient sur les API, permettant une collecte plus exhaustive des données composant les tweets. Twitter propose tout un écosystème d'API REST (pour Representational State Transfer) donnant accès à une grande variété de données (tweets, profils, hashtags, lieux, etc.), fournies dans le format JSON. La collecte de tweets peut être réalisée par mot-clef,

Licence CC-BY-SA

par zone géographique, ou encore par utilisateur, *a posteriori* (avec certaines limites, notamment pour la recherche par mot-clef, qui ne donne pas accès à des publications datant de plus d'une semaine) mais aussi en temps réel, à mesure qu'ils sont publiés. Si ces méthodes de collecte fournissent davantage de métadonnées que l'aspiration web, leur portée peut également être étendue par des requêtes récursives (portant par exemple sur les utilisateurs). Plusieurs outils clés en main, adossés à ces API, permettent de les exploiter dans des interfaces graphiques, à l'image de TCAT [Borra & Rieder, 2014], NodeXL et son module d'import de données Twitter [Smith *et al.*, 2010], Tweet Archivist, ou encore l'extension TwitterStreamingImporter pour Gephi. Des systèmes de collecte personnalisés peuvent également être mis en place à l'aide des nombreuses librairies logicielles existant pour un large choix de langages [Twitter, 2019f]. Enfin, des corpus complets de tweets issus des API peuvent être achetés, auprès de Twitter ou de fournisseurs tiers, notamment lorsque les données recherchées sont trop anciennes ou trop volumineuses pour être collectées directement.

La capture d'écran, enfin, consiste à enregistrer, sous la forme d'une image, ce qui est affiché sur l'écran d'un terminal donné. Elle permet ainsi de conserver une trace fidèle de la représentation graphique des tweets, mais sans leur dimension interactive. La capture peut être réalisée à partir des fonctionnalités directement intégrées dans la plupart des systèmes d'exploitation. Cependant, le processus peut être facilité par des outils en ligne tels que Screenshot Guru, ou des extensions de navigateur comme Twitter Screenshot pour Google Chrome, qui présentent notamment l'avantage de recadrer automatiquement les images produites. Différents outils en ligne de commande peuvent également être employés pour automatiser la capture d'un grand nombre de tweets à l'aide de scripts [Supriyo Biswas, 2018]. Le recours à des dispositifs tiers doit néanmoins se faire avec prudence : certains outils, comme Screenshot a Tweet, ne produisent pas de véritables *screenshots*, mais les construisent en intégrant des données de l'API dans une image de tweet vierge. Si le modèle utilisé ne correspond pas (ou plus) exactement à l'interface de Twitter au moment de la capture, la capture perd alors son caractère de fidélité.

Licence CC-BY-SA

Tableau 3. Éléments capturés par différentes modalités de collecte des tweets

Modalité de collecte	Données			Structure		
	Visibles	Métadonnées	Contexte	Graphique	Interaction	Informatique
Capture d'écran	Oui (format image)	Non	Non	Oui (pour un terminal)	Non	Non
Aspiration web	Oui	Non	Oui (seulement visibles)	Oui (pour un terminal)	Oui (pour un terminal)	Non
Aspiration web récursive	Oui	Non	Oui (seulement visibles)	Oui (pour un terminal)	Oui (pour un terminal)	Non
Requêtes API	Oui (images exclues)	Oui	Non	Non	Non	Oui (JSON)
Requêtes API récursives	Oui	Oui	Oui	Non	Non	Oui (JSON)

Comme on peut le constater dans le tableau 3, aucune des modalités de collecte présentées ici ne permet de sauvegarder intégralement les données et les structures qui constituent un tweet. La combinaison de plusieurs méthodes (par exemple, capture d'écran et requêtes API récursives), appliquées sur plusieurs terminaux (par exemple, site web, application iOS et application Android), est nécessaire pour se rapprocher de l'exhaustivité – mais sans jamais pouvoir l'atteindre tant les terminaux, et donc les structures, sont nombreux et variés.

ÉVOLUTIONS TEMPORELLES DES TWEETS

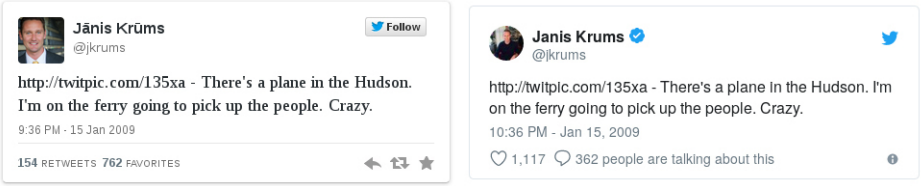
Si la collecte des tweets présente de nombreux écueils, la préservation des corpus ainsi constitués s'avère tout aussi complexe. En effet, cet archivage doit tenir compte de la nature changeante des tweets, dont les différents composants sont susceptibles de connaître plusieurs types d'évolution au cours du temps.

Certaines des données qui composent le tweet, tout d'abord, changent au gré des actions et des interactions des usagers. Le texte d'un tweet ne peut être modifié après sa publication, mais les retweets, les likes et les réponses peuvent s'y ajouter indéfiniment, parfois des années plus tard. Les données relatives aux profils utilisateurs (auteur ou mentionné), aux hashtags, aux aperçus de liens hypertextes, évoluent également avec le contexte dans lequel s'inscrivait

Licence CC-BY-SA

le tweet : l'auteur met à jour son image de profil ou sa biographie, un utilisateur mentionné change de pseudonyme, une page web citée voit son contenu modifié ou supprimé. Des tweets entiers (pouvant eux-mêmes être ou avoir des retweets ou des réponses) peuvent en outre disparaître lorsqu'ils sont supprimés par leur auteur, que celui-ci ferme son compte ou le paramètre comme « protégé » (dont les tweets ne sont visibles que par les utilisateurs autorisés). La grande majorité de ces changements qui affectent les constituants des tweets ne sont pas datés dans les systèmes d'information de Twitter.

Figure 2. Captures d'écran d'un même tweet tel qu'il apparaissait sur le site web de Twitter à deux dates différentes (4 novembre 2015 à gauche et 17 mai 2019 à droite) : < https://twitter.com/jkrums/status/1121915133 >



Les éléments de structure des tweets varient également au cours du temps, au fil des mises à jour de la plateforme Twitter. Depuis les débuts du site web, en 2006, la représentation graphique des tweets et les interactions qui y sont rattachées ont ainsi connu plusieurs changements majeurs (voir figure 2). On peut notamment mentionner l'intégration croissante de contenus multimédias (images, vidéos, aperçus des liens) et d'éléments cliquables (hashtags, mentions, indices boursiers) dans le corps des tweets, mais aussi des modifications dans leur ordre d'affichage, avec l'apparition des tweets sponsorisés, des systèmes non-chronologiques ou encore des fils (*threads*). Cette évolution progressive du système, ou *drift* [Salganik, 2018], se produit également – mais pas nécessairement de manière identique ou simultanée – dans les multiples applications (officielles ou développées par des tiers) permettant l'affichage de tweets sur divers terminaux. La représentation informatique des tweets qui est donnée à voir à travers les API connaît elle-même des changements, à mesure que des champs y sont ajoutés, renommés, dépréciés ou supprimés [Twitter, 2019d].

L'apparence du tweet dépend également d'un écosystème d'outils en constante évolution. Différentes versions d'un navigateur web ou d'un système d'exploitation (notamment pour les terminaux mobiles) peuvent affecter sa représentation graphique : pictogrammes, polices, émoticônes, espacements, etc. Par ailleurs, les contenus intégrés dans les tweets – images,

Licence CC-BY-SA

vidéos, etc. – reposent souvent sur des services tiers, dont les fonctionnalités peuvent changer ou disparaître au cours du temps. Par exemple, avant que Twitter ne permette directement la mise en ligne d'images, celles-ci étaient le plus souvent téléversées sur d'autres plateformes telles que TwittPic et yfrog, puis leurs URL incluses dans le texte du tweet. Selon les terminaux et les applications utilisés, un aperçu de l'image pouvait alors être automatiquement affiché. Si, à la fermeture de TwittPic, un accord a pu être trouvé avec Twitter pour que les images soient conservées [Twitpic, 2014], il n'en a pas été de même pour yfrog : l'ensemble des images a disparu, ne laissant que des liens morts dans quelques millions de tweets. De même, avant le déploiement de t.co par Twitter en 2010, la réduction des URL était assurée par des services tiers, TinyURL puis bit.ly. La validité de nombreux liens est par conséquent liée à la pérennité de ces services. Ainsi, la création de corpus de tweets complets et durables nécessite d'« archiver tout un écosystème numérique » [Clavert, 2018].

La variabilité temporelle des tweets se retrouve dans les documents numériques. Néanmoins, l'ampleur et la fréquence des modifications susceptibles d'affecter l'ensemble de leurs composants constituent une exception : « un document est défini par les éléments qui lui procurent une stabilité […] un certain nombre d'invariants qui régissent sa cohérence au sein des différentes formes qu'il peut revêtir » [Pédauque, 2006, p. 113-114]. Les invariants d'un tweet sont très limités : seuls le texte du tweet et certaines des métadonnées relatives à ses modalités de publication (identifiant unique, date et heure, langage, localisation, etc.) restent inchangés – tant que le tweet n'est pas supprimé. Tous les autres éléments doivent faire l'objet de procédures de collecte spécifiques afin de prendre en compte leurs valeurs successives.

CAPTURER L'ÉVOLUTION

La méthode la plus fréquemment utilisée pour enregistrer l'évolution d'un phénomène au cours du temps consiste à réaliser plusieurs collectes de données successives, qui pourront ensuite être comparées. Le projet Internet Archive, visant à archiver le web, suit cette approche : ses robots visitent et sauvegardent à intervalle régulier une grande quantité de pages web. Dans le cas des tweets, pour être complète, cette technique doit tenir compte de la nécessité – mentionnée précédemment – de combiner plusieurs méthodes de collecte.

La fréquence d'archivage constitue l'une des principales problématiques en matière de collecte récurrente. Une fréquence élevée permet de capturer avec finesse la dynamique d'évolution du tweet, de savoir précisément

Licence CC-BY-SA

quand une image a été modifiée, un like ajouté ou une réponse supprimée. Cependant, les opérations de collecte sont alors démultipliées, ainsi que les ressources nécessaires : temps de calcul, trafic réseau, espace de stockage. La fréquence d'archivage est alors un compromis à trouver entre la résolution temporelle souhaitée et les ressources disponibles. Des stratégies de collecte non-régulières peuvent constituer des solutions alternatives : la fréquence d'archivage d'un tweet donné peut être décroissante dans le temps, dans la mesure où une part importante des interactions (like, retweet, réponse, etc.) se produisent peu après la publication ; elle peut également être modulée en temps réel, en fonction des changements identifiés à chaque collecte [Saad & Gançarski, 2010]. Un stockage différentiel, n'enregistrant que les données ayant réellement changé d'une version à l'autre, peut également réduire le volume des corpus ainsi constitués.

Le processus de collecte lui-même peut également être affecté par certaines évolutions de la plateforme au cours du temps, notamment si elles concernent la structure des tweets. Lorsque de nouveaux éléments y sont ajoutés, retirés, renommés ou réorganisés, la représentation du tweet change, dans les pages HTML ou dans les données JSON fournies par les API. Les systèmes de capture doivent alors être adaptés de manière à continuer d'enregistrer l'ensemble des métadonnées disponibles. Par exemple, l'extension de la longueur des tweets de 140 à 280 caractères, en novembre 2017, a abouti à l'ajout de plusieurs nouveaux champs dans l'API [Twitter, 2019b], qui doivent à leur tour être intégrés dans les outils d'archivage. De même, des évolutions dans le fonctionnement profond de la plateforme, telles que les méthodes d'authentification ou les limites des API, peuvent avoir d'importants impacts sur le fonctionnement des outils. Il apparaît ainsi que l'instabilité des tweets dans le temps est suffisamment importante pour entraîner une instabilité du processus de collecte. Dès lors que l'on souhaite enregistrer l'évolution d'un corpus sur une période dépassant quelques semaines, il devient nécessaire d'assurer une veille technologique afin d'anticiper la survenue de ces changements et d'éviter des pertes de données ou une interruption pure et simple de la collecte.

Enfin, il apparaît essentiel de documenter ces évolutions de la plateforme pour guider l'analyse ultérieure des corpus qui auront été constitués. En effet, certains changements sont à l'origine de discontinuités dans les usages, qui ne pourraient être interprétées en l'absence d'information sur la date, la nature et l'impact des changements effectués. Ainsi, le 3 novembre 2015, Twitter remplaçait son bouton « mettre en favori » (*favorite* dans la version anglophone), associé à une étoile, par un bouton « j'aime » (like), en forme de cœur – qui n'est pas sans rappeler la fonction équivalente de Facebook [Twitter, 2015]. Ce nouveau bouton, à la sémantique très différente, ne recouvre que partiel-

Licence CC-BY-SA

lement les multiples usages qui s'étaient développés autour des favoris [Meier, Elsweiler & Wilson, 2014]. Pourtant, dans les métadonnées du tweet, les deux actions sont stockées dans un seul et même champ, comme si les « favoris » antérieurs à novembre 2015 avaient été transformés en « j'aime ». Si un tweet publié avant cette date présente un nombre donné de « j'aime », seule une capture de ce tweet réalisée précisément le 3 novembre 2015 permet de déterminer combien d'entre eux sont en réalité des « favoris », et de les analyser comme tels – à défaut, des estimations peuvent être effectuées sur la base des captures les plus proches de cette date. De la même manière, on pourrait souhaiter prendre en compte l'impact de l'évolution des affordances qui accompagnent les changements dans les interfaces graphiques – tels que le passage de l'injonction "*What are you doing*" à "*What is happening*" [Twitter, 2009].

Il apparaît ainsi que la mise en place d'une collecte de tweets capturant leur évolution au cours du temps est bien plus ardue que l'archivage périodique de pages web. La nature complexe des tweets et leur enchâssement dans la plateforme qui les héberge nécessitent un archivage plus fin et, par conséquent, moins robuste.

CONCLUSIONS ET DISCUSSIONS

Cet article a montré que la création d'archives des tweets est un défi majeur, pour les chercheurs qui étudient les usages des médias sociaux, mais aussi pour la société dans laquelle ils occupent une place centrale. Il apparaît néanmoins que les tweets, aussi simples qu'ils puissent sembler, sont caractérisés par la complexité et l'intrication des éléments qui les composent, ainsi que par des évolutions temporelles rapides, variées et parfois profondes. Afin de capturer l'ensemble des données permettant la reconstitution de ces éléments et de leurs formes successives, un processus de collecte complexe doit être mis en œuvre. Il s'agit notamment de combiner plusieurs méthodes de capture sur plusieurs terminaux, de manière répétée et à une fréquence adaptée, tout en documentant attentivement les changements susceptibles d'affecter la qualité, la complétude et l'interprétation des données. Lorsque les contraintes techniques, humaines ou économiques qui s'imposent à tout projet de collecte font obstacle à la mise en œuvre de l'intégralité de ce processus, des concessions (en termes de fréquence de capture, d'exhaustivité ou encore de plage temporelle) doivent être définies selon les usages prévus ou prévisibles des données. Les corpus constitués de cette manière sont nécessairement massifs et complexes, agrégeant des données textuelles et multimédias, dans des structures susceptibles de présenter des défauts et des discontinuités temporelles. Leur exploitation nécessite par conséquent le développement

Licence CC-BY-SA

de méthodes et d'outils d'analyse adaptés, permettant la re-présentation des tweets, mais aussi de leur contexte et de leur évolution temporelle.

Par ailleurs, il apparaît en plusieurs points de cet article que la collecte des tweets gagne à prendre en compte les points communs de ces derniers avec les documents numériques. Comme eux, les tweets peuvent être décrits comme la combinaison de données et de structures, n'existent que dans un environnement technique précis, voient leurs représentations se reconfigurer selon les supports, et connaissent une évolution temporelle. Ils ne se démarquent que par le degré extrême qu'atteignent ces différentes caractéristiques : les tweets sont particulièrement instables dans le temps, particulièrement dépendants de leur plateforme et des terminaux d'affichage, et se constituent d'un enchevêtrement particulièrement complexe de données et de structures. Il semble par conséquent pertinent de considérer les tweets comme un type de documents numériques, possédant certes des spécificités, mais également de nombreuses similarités avec les autres. En particulier, leur archivage est guidé par le même impératif de re-présentation et de remise en contexte : « Il s'agit alors d'apporter toutes les métadonnées indispensables à la reconstruction à la volée de documents et toute la traçabilité de son cycle » [Salaün, 2007]. À ce titre, le processus de collecte des tweets décrit ici constitue une forme de redocumentarisation.

Enfin, plusieurs éléments doivent être relevés quant à la conformité des techniques de collecte présentées avec les conditions d'utilisation « Développeurs » de Twitter. En premier lieu, la redistribution des corpus collectés sur la plateforme est strictement encadrée [Twitter, 2019c]. L'envoi de données complètes (telles que fournies par les API) est limité à 50 000 tweets par jour et par destinataire, et exclut la possibilité de mettre ces données à disposition par téléchargement. Pour l'échange de corpus plus importants, Twitter recommande de n'envoyer que les identifiants des tweets – une limite de 1 500 000 identifiants tous les 30 jours est définie, mais les chercheurs peuvent en être exemptés – laissant le soin au destinataire de collecter à nouveau les données associées. Cette restriction fait obstacle à tout effort de collecte tenant compte de l'évolution des tweets au cours du temps, puisque seule une version récente du tweet pourrait être partagée. Par ailleurs, les conditions d'utilisation indiquent clairement que les tweets ultérieurement supprimés, modifiés ou protégés devraient être rapidement retirés des jeux de donnés [Twitter, 2019a]. Cette mesure, bien compréhensible du point de vue des utilisateurs, s'avère difficile à mettre en œuvre et, à nouveau, incompatible avec l'étude de l'évolution des tweets au cours du temps. Les entraves légales que constituent ces textes contractuels – au même titre que les contraintes techniques relatives à l'accès aux API [Rieder, 2018] – doivent faire l'objet

Licence CC-BY-SA

d'une réflexion critique. Il s'agit de questionner la capacité qu'ont et devraient avoir les chercheurs – mais aussi la société – à documenter et étudier des dispositifs aussi importants que les médias sociaux.

BIBLIOGRAPHIE

Alexa Internet. (2019). Twitter Competitive Analysis, Marketing Mix and Traffic. < https://www.alexa.com/siteinfo/twitter.com >.

Blumenthal, K.-R. (2019). Archiving Twitter feeds. < https://support.archive-it.org/hc/en-us/articles/208333743-Archiving-Twitter-feeds >.

Borra, E., & Rieder, B. (2014). Programmed Method : Developing a Toolset for Capturing and Analyzing Tweets. *Aslib Journal of Information Management*, 66 (3), 262-278.

Bruns, A. (2018). The Library of Congress Twitter Archive : A Failure of Historic Proportions. < https://medium.com/dmrc-at-large/the-library-of-congress-twitter-archive-a-failure-of-historic-proportions-6dc1c3bc9e2c >.

Cervulle, M., & Pailler, F. (2014). #mariagepourtous : Twitter et la politique affective des hashtags. *Revue française des sciences de l'information et de la communication*, (4). < http://journals.openedition.org/rfsic/717 >.

Clavert, F. (2018). Sources en flux. Collecter, analyser, archiver, pérenniser. In A. Francois, A. Roekens, V. Fillieux, & C. Derauw, *Pérenniser l'éphémère. Archivage et médias sociaux*. Louvain-la-Neuve, Belgique : Academia, p. 23-44.

Durupt, F. (2019). Pourquoi Twitter bannit-il les « publicités politiques » ? < https://www.liberation.fr/evenements-libe/2019/10/31/pourquoi-twitter-bannit-il-les-publicites-politiques_1760777 >.

INA. (2019). Comptes Twitter liés à l'audiovisuel français. < https://www.data.gouv.fr/fr/datasets/comptes-twitter-lies-a-laudiovisuel-français/ >.

INA. (s. d.). Dépôt légal radio, télé et web. < https://institut.ina.fr/institut/statut-missions/depot-legal-radio-tele-et-web >.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite ! The challenges and opportunities of Social Media. *Business horizons*, 53 (1), 59-68.

Library of Congress. (2017). Update on the Twitter Archive at the Library of Congress. < https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf >.

Meier, F., Elsweiler, D. C., & Wilson, M. L. (2014). More than liking and bookmarking Towards understanding twitter favouriting behaviour. *Eighth International AAAI Conference on Weblogs and Social Media*.

Mercier, A. (2015). Twitter, espace politique, espace polémique. *Les cahiers du numérique*, 11 (4), 145-168.

Molina, B. (2017). Twitter overcounted active users since 2014, shares surge on profit hopes. USA Today. < https://eu.usatoday.com/story/tech/news/2017/10/26/twitter-overcounted-active-users-since-2014-shares-surge/801968001/ >.

Licence CC-BY-SA

Musiani, F., Paloque-Bergès, C., Schafer, V., & Thierry, B. G. (2019). *Qu'est-ce qu'une archive du web ?* Marseille, France : OpenEdition Press.

Musiani, F., & Schafer, V. (2016). Patrimoine et patrimonialisation numériques. *RESET. Recherches en sciences sociales sur Internet*, (6).

Ott, B. L. (2017). The age of Twitter : Donald J. Trump and the politics of debasement. *Critical studies in media communication*, 34 (1), 59-68.

Paloque-Bergès, C. (2016). Les sources nativement numériques pour les sciences humaines et sociales. *Histoire@Politique*, 30 (3), 221-244.

Pédauque, R. T. (2006). *Le Document à la lumière du numérique : forme, texte, médium : comprendre le rôle du document numérique dans l'émergence d'une nouvelle modernité*. Caen, France : C & F Éditions.

Potts, L., Seitzinger, J., Jones, D., & Harrison, A. (2011). Tweeting disaster : Hashtag constructions and collisions. *Proceedings of the 29th ACM international conference on Design of communication*, 235-240. ACM.

Rieder, B. (2018). Facebook's app review and how independent research just got a lot harder. < http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/ >.

Rogers, R., Brügger, N., & Milligan, I. (2018). Periodizing web archiving : Biographical, event-based, national and autobiographical traditions. In *The SAGE Handbook of Web History* (p. 42).

Saad, M. B., & Gançarski, S. (2010). Using visual pages analysis for optimizing web archiving. *Proceedings of the 2010 EDBT/ICDT Workshops*, 43. ACM.

Salaün, J.-M. (2007). La redocumentarisation, un défi pour les sciences de l'information. *Études de communication. Langages, information, médiations*, (30), 13-23.

Salganik, M. (2018). *Bit by Bit, Social reseach in the digital age*. Princeton, États-Unis : Princeton University Press.

Smith, M., Milic-Frayling, N., Shneiderman, B., Mendes Rodrigues, E., Leskovec, J., & Dunne, C. (2010). *NodeXL : a free and open network overview, discovery and exploration add-in for Excel 2007/2010*. Social Media Research Foundation.

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes : Misinformation on twitter after the 2013 Boston marathon bombing. *iConference 2014 Proceedings*, 654-662.

Supriyo Biswas. (2018). How to Take Screenshots of Webpages from the Command Line. < https://www.booleanworld.com/take-screenshots-webpages-command-line/ >.

Twitpic. (2014). Twitpic's Future. < https://web.archive.org/web/20141027024335/http://blog.twitpic.com/2014/10/twitpics-future/ >.

Twitter. (2009). What's Happening? < https://blog.twitter.com/en_us/a/2009/whats-happening.html >.

Twitter. (2015). Hearts on Twitter. < https://blog.twitter.com/official/en_us/a/2015/hearts-on-twitter.html >.

Licence CC-BY-SA

Twitter. (2019a). Developer Policy. < https://developer.twitter.com/en/developer-terms/policy#c-respect-users-control-and-privacy >.

Twitter. (2019b). Extended Tweets. < https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json#extendedtweet >.

Twitter. (2019c). More about restricted uses of the Twitter APIs. < https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases >.

Twitter. (2019d). Tweet metadata timeline. < https://developer.twitter.com/en/docs/tweets/data-dictionary/guides/tweet-timeline >.

Twitter. (2019e). Tweet Objects. < https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object >.

Twitter. (2019f). *Twitter Libraries*. < https://developer.twitter.com/en/docs/developer-utilities/twitter-libraries.html >.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events : What twitter may contribute to situational awareness. *Proceedings of the SIGCHI conference on human factors in computing systems*, 1079-1088. ACM.

Licence CC-BY-SA