*La mise à disposition des archives de Twitter
par la Library of Congress
Aspects méthodologiques et difficultés de process*

Jean-Daniel ZELLER

Le 1er avril 2010, la Bibliothèque du Congrès (LOC) et Twitter ont annoncé le projet commun d’archiver tous les tweets publics de la plateforme Twitter. En 2012, la LOC a annoncé avoir reçu les premiers versements historiques couvrant la période 2006-2008 (env. 133 Terabytes) puis des versements réguliers en constant accroissement (Zimmer, 2015).

Même si la LOC est une des institutions expérimentée dans le traitement des documents électroniques, les objets de nature particulière que sont les tweets posent des défis inédits qu’elle n’a pas encore entièrement résolut à ce jour, ce qui rend ces fonds encore inaccessibles aux chercheurs, plus de 5 ans après l’annonce initiale.

Ces obstacles, de nature techniques ou archivistiques, sont les suivants :

1. La grande volumétrie,
2. L’abondance des métadonnées
3. La question de la protection de la vie privée
4. La question de la suppression rétrospective des tweets
5. Les chercheurs « bona fide »
6. La problématique de l’archivage des plateformes mondialisées
7. Twitter est-t-il un réseau social ou un simple média ?
8. Archiver Twitter a-t-il une utilité ?

1. La grande volumétrie

La LOC, en acceptant le versement des archives de Twitter a doublé d’un seul coup le volume de ses archives numériques, au moment des deux versements initiaux des tweets de la période 2006 2008. Les versements ultérieurs par tranches annuelles ont été en constante augmentation. Cette évolution n’est pas documentée par la LOC ni rétrospectivement par Twitter, mais si l’on compare les chiffres accessibles de 2011 et de 2015 on a le tableau suivant :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **tweets/jour** | **Go/jour(tw x 140)** | **Go/an** | **To/an** | **To/anavec MD** |
| 2011 | 65 millions | 9.1 | 3’321,5 | 3 | **58** |
| 2015 | 499 millions | 69.9 | 36’463,5  | 36 | **576** |
| *Multiple* | *7,7* | *7,7* | *12* | *12* | *16* |

*Tableau 1 : Evolution de la volumétrie de Twitter (2011-2015)*

**2.°L’abondance des métadonnées**

Une des particularités des archives de Twitter est que les métadonnées associées aux tweets sont beaucoup plus volumineuses que les tweets eux-mêmes. Dans son article de 2014 titré “*In a Single Tweet, as Many Pieces of Metadata as There Are Characters*” (que l’on peut traduire par « Dans un seul tweet, autant de champs de métadonnées qu’il y a de caractères »), Elizabeth Dwoskin indique qu’il y plus de d’une centaine de champs de métadonnées associées à chaque tweet (Dwoskin, 2015). Elle fait référence à un billet de Raffi Krikorian, ingénieur chez Twitter, qui en a donné le détail en 2010 (Krikorian, 2010).

Cela a pour conséquence que les volumes indiqués précédemment deviennent 16 fois plus important en tenant compte q’un tweet de 140 caractères génère environs 2250 caractères de métadonnées (voir la dernière colonne du tableau ci-dessus).

**3.°La question de la protection de la vie privée**

La convention liant la LOC à Twitter a fixé un délai de 6 mois pour la communication des tweets archivés. Suite à cela, Twitter a été accusé d’utiliser abusivement des données privées sans consentement. En effet, cette opération d’archivage ne fait pas partie des clauses contractuelles de l’abonnement à Twitter et la majorité des utilisateurs n’utilisent pas l’option « privé » pour leurs messages, ce qui les excluraient de cet archivage de masse. Comme il n’y a pas eu divulgation à ce jour, il n’y a pas encore de procédure en cours mais il est probable que des actions en justice puissent êtres entreprises à terme, à moins que la LOC ne trouve une méthode d’anonymisation suffisamment robuste. Même si certaines études récentes mobilisent des méthodes qualitatives (entre autres exemples, Bosch, 2017), la plupart des chercheurs recourent habituellement au traitement de masse. On peut dès lors penser que cette contrainte d’anonymisation ne suscitera guère de résistances.

**4.La question des suppressions rétrospectives de tweets**

Dans la même veine, des auteurs ont supprimé certains tweets sujets à controverse, suppression qui n’est pas effective dans les tweets archivés, qui ne tiennent pas compte du « remord » de leurs auteurs.

Cette question a été partiellement abordée dans le cas de « repentir » d’hommes politiques. Twitter, qui dans un premier temps avait fait fermer les sites « Politwoops »[[1]](#footnote-1) qui traçaient ce type de pratique, parce qu’ils semblaient contraires à ses conditions générales, est finalement revenu en arrière en décembre 2015. Cela a fait l’objet d’un accord tripartite entre Twitter, la Sunligth Foundation et l’Open State Foundation, ces deux dernières représentant les milieux défendant la transparence politique.

La question de savoir si ce qui s’applique aux hommes politiques doit s’appliquer au twitteur lambda reste ouverte. Elle renvoie également à la question de savoir si Twitter est réellement un média social ou un simple média, ce que nous développerons plus loin.

**5.Les chercheurs « bona fide »**

Une autre clause de l’accord de versement stipule que les archives seront communiquées à des chercheurs de « bonne foi » ce qui exclut un accès libre à tout public, en contradiction avec l’esprit des archives publiques auquel devrait logiquement adhérer la LOC en tant qu’institution publique. Il n’y a pas actuellement d’information de la LOC sur le processus de désignation et/ou d’habilitation de ces chercheurs. Cette clause a agité les milieux des sciences de l’information américains qui ont rappelés de manière constante les principes archivistiques du libre accès mais aussi du délai de protection nécessaire à la vie privée (voir : Chen, 2014 ; Lipton 2010 et Vieweg, 2010 ; cité dans Zimmer, 2015). Dans ce dernier cas il est ironique de penser que le délai avait été fixé à six mois, ce qui est extrêmement court eu égard aux délais de protection habituellement pratiqués en la matière (plutôt 30 à 100 ans) mais qu’effectivement les problèmes techniques allongent pratiquement ce délai puisque qu’aucun tweet n’est encore accessible par cette voie (ce qui fait déjà près de 10 ans pour les premiers tweets de 2006).

**6. La problématique de l’archivage des plateformes mondialisées**

L’archivage des plateformes mondialisées, comme celui du web en général, interpelle une pratique archivistique assez uniformément appliquée de longue date qui est celui de la territorialité de la provenance. Dans le cadre du web, le partage des tâches entre les bibliothèques nationales qui sont les principaux acteurs dans ce domaine n’a pas empêché un acteur privé (Internet Archives) d’offrir une collecte mondialisée. Par ailleurs la coexistence d’identifiants géographiques (.fr, .ch, .be) et d’identifiants fonctionnels (.org, .com, .edu) posent nombre de questions pour la programmation des robots moissonneurs de pages web. On peut lire à ce sujet un récent billet de la nouvelle présidente de l’ International Internet Preservation Consortium (IIPC)[[2]](#footnote-2) qui expose le problème de façon détaillée, sans y apporter de réponse[[3]](#footnote-3).

Dans le cadre des réseaux sociaux, il apparaît de plus en plus évident que la pratique archivistique ne relève pas de l’archivage au sens strict, c'est-à-dire le transfert des documents/données tant sur le plan physique que sur le plan de la responsabilité de conservation, mais plutôt une « copie » (idéalement anonymisée) de la masse (une sorte de vrac numérique) alors que les données de base restent (ad infinitum ?) sur les serveurs de la plateforme, où une archive « ad personam » reste accessible au possesseur du compte lui-même. Cependant, si ces données sont « affichables », la plupart du temps elles ne sont pas « traitables ». A titre d’exemple les adresses en notes[[4]](#footnote-4) sont celles des « archives » des comptes de réseaux sociaux de l’auteur et ne sont en principe pas accessibles par d’autres que lui.

On comprend bien l’intérêt technique de l’opération pour les propriétaires des plateformes qui s’épargnent ainsi la prise en charge de temps machine pour des extractions ou des tris complexes, tout en se réservant le droit de commercialiser ce service auprès des intéressés, dans le cas de Twitter par l’entremise du la société GNIP[[5]](#footnote-5).

**7.Twitter est-il un réseau social ou un simple média ?**

Cette question a été posée par des chercheurs coréens dès 2010 (Kwak et al., 2010). Utilisant les API alors mis à disposition par Twitter, ils ont collectés la quasi totalité des tweets émis entre le 6 et le 9 juillet 2009, ce qui leur a permis de recenser 41,7 millions de profils d’utilisateur, représentant 1,47 milliard de liens (suiveur/suivi) et ayant produit 106 millions de tweets.

Dans leur conclusion, les auteurs notent que l’écosystème de Twitter ne montre pas une distribution suivant une loi de puissances ; il possède un diamètre d’efficacité court, et une faible réciprocité (entre suiveurs et suivis). Ces trois éléments sont l’indicateur que Twitter ne semble pas suivre les critères standards d’un réseau social humain. Ceci est illustré par la figure ci-dessous, issue de leur article, qui modélise les liens entre les différents suiveurs du hashtag « «#Air France flight » utilisé lors de la chute catastrophique d’un avion de cette compagnie. La partie de gauche montre la partie des twitteurs « en réseau » et l’importante partie de droite montre une part bien plus grande de twitteurs « hors réseau », ce qui accrédite l’hypothèse des auteurs.

*Figure 1 : Arbre des retweets avec le tag « Air France flight » (d’après Kwak et al. (2010))*

D’autres observateurs francophones des réseaux sociaux défendent la même hypothèse, sans l’étayer de manière aussi formelle (Cavazza, 2015 ; Ertzscheid, 2012).

**8.Archiver Twitter a-t-il une utilité ?**

Au moment ou la LOC a annoncé l’archivage de Twitter, certains milieux académiques nord-américains se sont inscrit en faux contre cette opération au prétexte que ce contenu était tellement pauvre qu’il ne représentait aucun intérêt à long terme. Pour faire court, la comparaison avancée était que l’on n’avait jamais trouvé de l’intérêt à archiver les discussions du type « café du commerce ».

A contrario, l’argument avancé par la LOC est que l’archivage de Twitter (ou d’autres réseaux sociaux) permettait justement de comprendre certains phénomènes sociaux jusqu’alors intraçables. Malheureusement, la lecture de la bibliographie sur les recherches à propos des réseaux sociaux établie par danah boyd[[6]](#footnote-6) semble laisser supposer, sous réserve d’une analyse plus approfondie, que les recherches sur ces corpus ne mettent à jour que des truismes. L’étude de Kwak et al. (2011), en mettant en évidence le rôle des retweets et en excluant Twitter du monde des réseaux sociaux stricto sensu semblent néanmoins monter qu’en posant de bonnes questions on peut trouver des réponses pertinentes.

Je reprends ici l’argumentaire d’Olivier Ertzscheid (2012) qui me paraît exprimer une analyse correcte du choix de la LOC :

*« Il est 2 raisons d'être à la science de l'archivistique : l'obligation et le choix. L'obligation d'archiver des supports arrivés à la fin de leur cycle de vie et le choix des "items documentaires" qui, parmi l'ensemble de ces supports en fin de cycle, sont dignes d'intérêt ou recouvrent une certaine "valeur".*

***Les 3 (vraies?) raisons.*** *Il y a - à mon avis - trois raisons profondes à l'archivage de la totalité de Twitter ;*

***Primo : la simplicité.*** *Il est beaucoup plus simple de prendre, en vrac, la totalité de l'archive twitterienne que de l'investiguer pour y effectuer un travail de sélection en amont.*

***Deuxio : le graphe social.*** *L'archive de Twitter ainsi constituée permettra  - à des chercheurs, sociologues, historiens - de reconstituer le graphe relationnel de chaque individu choisi.*

***Tertio : l'archive elle-même.*** *Si la LoC s'intéresse à Twitter, un service […] dont rien ne permet de dire qu'il sera ou non pérenne sous sa forme actuelle, si la LoC s'intéresse à Twitter, à la totalité de Twitter, c'est parce dans cette totalité transparaît l'essence même de l'objet documentaire idéal.*

La première raison semble être pertinente à postériori puisque la LOC a met du temps à développer des outils d’analyse (pas encore disponibles). La deuxième raison semble trouver sa justification dans les travaux de Kwak et al. (2011) ; En effet, sans une collecte exhaustive dans un temps T, il aurait été impossible à ces chercheurs de prouver que Twitter est un média plutôt qu’un réseau social stricto sensu Quant à la troisième elle représente certainement l’idéal du chercheur qui peut avoir à sa disposition un corpus idéalement exhaustif, alors que l’archiviste continuera probablement à défendre l’option d’un tri après évaluation.

**9.Twitter survivra-t-il ou deviendra-t-il le premier « fonds clos » numérique mondial ?**

Deux annonces récentes de Twitter pourraient mettre en péril cette plateforme. L’une porte sur la manière de gérer l’accès aux tweets :

Si cela se confirmait, ce serait un petit pas dans l'histoire des réseaux sociaux, mais un grand chamboulement pour les plus de 300 millions d'utilisateurs actifs que compte Twitter. [Selon Alex Kantrowitz du site *BuzzFeed News*](http://www.buzzfeed.com/alexkantrowitz/twitter-to-introduce-algorithmic-timeline-as-soon-as-next-we), le réseau social serait sur le point de proposer une « timeline algorithmique » à ses utilisateurs, sur base de leurs centres d’intérêts supposés. « *Cette ‘timeline’ réorganisera les tweets* [visibles par les utilisateurs] *à partir de ce que l’algorithme mis en place par Twitter pense que les gens ont le plus envie de voir* »[[7]](#footnote-7).

L’autre porte sur la limite « sacrée » des 140 caractères :

« *Twitter envisage de ne plus limiter à 140 caractères la taille des messages postés par les internautes. Le nombre de 10’000 signes circule. Aux yeux de nombreux utilisateurs, cette évolution dénaturerait le réseau social.*

*Twitter travaille sur une fonctionnalité qui permettrait de poster des messages plus longs. Limités à 140 caractères jusqu'à présent, les tweets pourraient bientôt courir sur 10’000 signes*»[[8]](#footnote-8).

Outre que les problèmes de volumétrie cités plus haut ne feront donc que s’accroître, ces deux options, qui feraient évoluer les fonctionnalités qui sont la marque de fabrique de la plateforme, pourraient, au dire des commentateurs, causer la disparition même de ce service. On ne pourrait alors que féliciter la LOC d’avoir eu la bonne intuition en préservant un corpus voué autrement à la disparition.

**Conclusion**

A la suite de ce parcourt quelque peu chaotique, on perçoit que l’archivage des réseaux sociaux est loin d’être une sinécure. Ceci d’autant plus que la plateforme Twitter représente un « petit » réseau social ; on estime par exemple que Facebook représente un volume d’utilisateurs près de cinq fois plus important (1320 millions contre 284 millions en 2014)[[9]](#footnote-9). A défaut de pouvoir accéder aux archives complètes de Twitter, les chercheurs en sont actuellement réduit à accéder à des vues (très) partielles et sur la base d’un échantillonnage fort aléatoire, comme le démontrent d’autres contributions à cet ouvrage[[10]](#footnote-10). Je doute que la recherche menée par Kwak et al. en 2010 puisse être encore réalisable à l’identique en 2016, ce qui serait néanmoins fort instructif.

Ceci nous amène à une hypothèse de travail qui est que, pour l’instant, il nous manque une algorithmique pertinente pour analyser de tel corpus. Comme nous serons néanmoins appelés à archiver de plus en plus de tels « vrac de données »[[11]](#footnote-11), que cela soit des réseaux sociaux ou d’autres fonds né-numériques, je plaide ici pour que la communauté archivistique commence à réfléchir à une **« algorithmique archivistique »** qui aille au-delà des mirages annoncés par les thuriféraires des mégadonnées.

**Bibliographie**

# *Tous les liens internet ont été consultés le 15 mars 2016*

# Bermès E. (2016), IIPC 2016 – how to collaborate ?, Figoblog, 25/04/2016 ; https://figoblog.org/2016/04/25/iipc-2016/

Bosch, T. (2017). Twitter activism and youth in South Africa: the case of #RhodesMustFall. *Information Communication and Society*, *20*(2), 221–232. https://doi.org/10.1080/1369118X.2016.1162829

boyd d. (2013). **Bibliography of research on Twitter & microblogging**
http://www.danah.org/researchBibs/twitter.php

# Cavazza F. (2015). Facebook est un portail, Twitter est un média. http://www.fredcavazza.net/2015/04/02/facebook-est-un-portail-twitter-est-un-media/

Chen A. (2014). **The laborers who keep dick pics and beheadings out of your Facebook feed,** Wired (23 October). http://www.wired.com/2014/10/content-moderation/.

Dwoskin E. (2014). **In a single tweet, as many pieces of metadata as there are characters,** WSJ Blogs — Digits (6 June), http://blogs.wsj.com/digits/2014/06/06/in-a-single-tweet-as-many-pieces-of-metadata-as-there-are-characters/

Ertzscheid O. (2012), **Twitter : un patrimoine superflu(x) ?**
http://affordance.typepad.com/mon\_weblog/2012/06/agenda-loeuvre-et-larchive.html

Guillaud H. (2016), **Numérique : la taille cet impensé**, InternetActu.net
http://www.internetactu.net/2016/05/04/numerique-la-taille-cet-impense/

Krikorian R. (2010), **Map of a Twitter Status Object** http://online.wsj.com/public/resources/documents/TweetMetadata.pdf

Kwak H., Lee C., Park H., and Moon S. (2010). **What is Twitter, a Social Network or a News Media?** *Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA)*
http://www.eecs.wsu.edu/~assefaw/CptS580-06/papers/2010-www-twitter.pdf

Eric Lipton, 2010. **Don’t look, don’t read: Government warns its workers away from WikiLeaks documents**, New York Times (4 December). http://www.nytimes.com/2010/12/05/world/05restrict.html.

Vieweg S. (2010). **The ethics of Twitter research**, Revisiting Research Ethics in the Facebook Era: Challenges in Emerging CSCW Research.

Zimmer M. (2015), **The Twitter Archives at the Library of Congress: Challenges for information practice and information Policy.** First Monday, 20,7 6 July 2015. http://firstmonday.org/ojs/index.php/fm/article/view/5619/4653

1. https://twitter.com/politwoops?lang=fr [↑](#footnote-ref-1)
2. International Internet Preservation Consortium (http://www.netpreserve.org/) [↑](#footnote-ref-2)
3. Voir Bermès E. (2016). [↑](#footnote-ref-3)
4. Facebook (https://www.facebook.com/jeandaniel.zeller/allactivity)
Twitter (https://analytics.twitter.com/user/JanusHelveticus/home)
Linkedin (https://www.linkedin.com/vsearch/f?adv=true&trk=federated\_advs)
Wordpress (<https://regarddejanus.wordpress.com/wp-admin/index.php> )
Lien API de recherche Twitter (<https://dev.twitter.com/rest/public/search>) [↑](#footnote-ref-4)
5. https://gnip.com/sources/twitter/ [↑](#footnote-ref-5)
6. Voir boyd d. (2013). [↑](#footnote-ref-6)
7. Brabant, J. (2016) « Twitter, un nouvel algorithme suicidaire ». *ArretsurImages*, <http://www.arretsurimages.net/breves/2016-02-06/Twitter-un-nouvel-algorithme-suicidaire-id19654> [↑](#footnote-ref-7)
8. M.S. (2016) « Twitter : bientôt 10 000 signes par message ? Autant l’appeler Facebook ! ». *L’express Tendances*, http://www.lexpress.fr/tendances/produit-high-tech/twitter-bientot-10-000-signes-par-message-autant-l-appeler-facebook\_1751019.html [↑](#footnote-ref-8)
9. Voir sur cette thématique Guillaud H. (2016). [↑](#footnote-ref-9)
10. Voir les contributions d’Alexandre Turgeon et d’Hubert Naets dans ce même ouvrage. [↑](#footnote-ref-10)
11. Cette thématique est d’actualité et a fait l’objet de communication lors du récent Forum des archivistes français à Troyes (http://forum2016.archivistes.org/blog/2016/01/20/atelier-mecano-traiter-un-vrac-numerique-les-outils-de-pre-versement/). [↑](#footnote-ref-11)