Table des auteurs

Frédéric CLAVERT, Maître assistant, Université de Lausanne.

Hervé COLIN, Webmaster fondateur du Wikipasdecalais (association et site Internet), Administrateur du Comité d'histoire du Haut-Pays.

Marie CHOULEUR, Cheffe du Service du dépôt légal numérique, Direction des services et des réseaux, Département du dépôt légal de la Bibliothèque nationale de France.

Florian DELABIE, Chargé de projet, Secteur Gestion de l'Information et Archives de la RTBF.

Aurore FRANÇOIS, Professeure et Archiviste de l'Université catholique de Louvain.

Alexandre GARCIA, Chef de projet, Division des archives et de la gestion de l'information du Comité international de la Croix-Rouge.

Odile GAULTIER-VOITURIEZ, Responsable de la documentation et des archives, Centre de recherches politiques de Sciences Po, Paris.

Annick LE FOLLIC, Chargée de collections numériques, responsable de la collecte, Département du dépôt légal de la Bibliothèque nationale de France.

Sophie LEGER, Présidente de Wikipasdecalais, Animatrice culturelle du Comité d'Histoire du Haut-Pays.

Lola MIRABAIL, Conservatrice des bibliothèques actuellement en poste au sein de la bibliothèque universitaire de Paris VIII comme responsable du département des services au public.

Hubert NAETS, Linguiste informaticien, CENTAL, Université catholique de Louvain.

Ivan PACHEKA, Administrateur du site Wikipasdecalais, Chef du service des archives contemporaines aux Archives départementales du Pas-de-Calais.

Patrick PECCATTE, Chercheur associé au Laboratoire d'Histoire visuelle contemporaine, EHESS.

Ulrike Lune RIBONI, Centre d'études sur les médias, les technologies et l'internationalisation, Paris VIII.

Anne ROEKENS, Professeure, Université de Namur.

Alexandre TURGEON, Chercheur postdoctoral, Université d'Ottawa.

Lidia UZIEL, Directrice du Département des Langues occidentales et conservatrice pour l'Europe de l'Ouest à la Bibliothèque de Harvard.

Thierry VEDEL, Chercheur, Centre de recherches politiques de Sciences Po, Paris.

Jean-Daniel ZELLER, Consultant indépendant en archivistique et gestion des documents d'activité.

Sources en flux. Collecter, analyser, archiver, pérenniser

Enseignements d'une collecte de tweets sur le Centenaire de la Grande Guerre

Frédéric CLAVERT

1. Introduction : aux sources d'une recherche

Qu'est-ce qu'une source primaire en flux ? Nous allons ici nous pencher particulièrement sur un réseau social numérique, Twitter, édité par la société californienne *Twitter, Inc.* Les traces (Boullier, 2015) ou données produites par cette plateforme du web font dans ce chapitre l'objet d'une tentative d'analyse en tant que sources primaires historiennes en flux. Elles sont vues comme un flot constant d'informations, dont l'intensité peut changer, mais qui est toujours dans une situation de fort contraste avec les archives, sources primaires habituelles de l'historien.ne, qui, elles, sont figées.

Twitter est un réseau social numérique, qui permet aux utilisateurs qui y ont ouvert un compte de publier des messages de 140 caractères (*tweets*) correspondant à la longueur initiale des textos sur téléphone portable, à laquelle Twitter a supprimé vingt caractères, réservés pour le nom d'utilisateur. Sur la page d'accueil de Twitter, une fois l'utilisateur connecté, se déroulent le flux des tweets émis par les comptes qui sont suivis par cet utilisateur (on parle de *timeline*), tout comme ses tweets apparaîtront dans la *timeline* des comptes qui suivent (*followers*) cet utilisateur. Il est possible de mentionner un autre utilisateur dans un tweet : le nom de compte de cet utilisateur apparaîtra alors dans le tweet précédé d'un « @ ». La mention sert à signaler un fait à ou à discuter avec un ou plusieurs autre(s) utilisateur(s). Il est également possible de retweeter tel quel ou avec commentaire le tweet d'un autre utilisateur (*retweet*). On peut insister sur un mot en le faisant précéder d'un croisillon (« # ») : on parle alors de *hashtag*. Le *hashtag*, ou mot-dièse, correspond à des usages très variés : insister sur un

mot, ironiser sur un concept, rejoindre une discussion impliquant de nombreux utilisateurs, etc. Enfin, on peut inclure une adresse URL pointant vers un site web permettant de détailler ainsi une information dans le cas, très fréquent, où 140 caractères ne suffisent pas.

L'usage de Twitter produit, donc, des traces (les tweets notamment), que l'on peut collecter dans la mesure où l'on respecte les conditions générales d'utilisation du service. Cette collecte peut prendre différents aspects. En mettant en avant un projet de recherche, fondé sur une base de données regroupant des tweets liés au Centenaire de la Première Guerre mondiale, nous allons ici développer certains enjeux auxquels l'historien.ne doit faire face lorsqu'il.elle utilise ce type de données. Pour renforcer cette approche, nous allons, dans cette introduction, détailler le chemin qui nous a amenés à notre recherche sur le Centenaire. L'une des étapes les plus importantes a été l'apprentissage d'un savoir-faire.

Apprentissage d'un savoir-faire

En 2009, lorsque mon employeur de l'époque, le *Centre Virtuel de la Connaissance sur l'Europe*[6], m'envoie aux États-Unis suivre la conférence *Digital Humanities 2009*[7], j'avais déjà créé un compte Twitter depuis environ un an, sans comprendre à quoi servait ce réseau social en ligne. Le cycle de conférence *Digital Humanities*, depuis 2006, réunit chaque année de nombreux acteurs de tout un champ de recherche. Comme de nombreuses conférences de taille importante, l'organisation repose sur des sessions menées en parallèle. Devant l'impossibilité de choisir les meilleures sessions, tant elles étaient riches, les participants utilisaient Twitter pour communiquer d'une session à l'autre, pour faire un résumé en direct des présentations, pour suivre, parfois, d'autres sessions lorsque l'on s'aperçoit que son choix n'a pas été pertinent. Ce fut également une *follow party* : chacun suivant ou étant désormais suivi par d'autres chercheurs, bibliothécaires, archivistes ou ingénieurs. En 2009, Twitter devint ainsi un élément fondamental des Humanités numériques vues comme une communauté de pratiques.

La partie francophone de cette communauté n'a pas immédiatement suivi. Vers 2012 cependant, un certain nombre de conférences ont joué le rôle, au niveau francophone, de *Digital Humanities 2009* et notamment les conférences *Digital Humanities*

---

[6] http://www.cvce.eu/.

[7] http://mith.umd.edu/dh09/.

Luxembourg 2012 (DHLU2012)[8], THATCamp Luxembourg / Trier[9] et THATCamp Paris 2012[10].

DHLU2012 était organisé par le Centre Virtuel de la Connaissance de l'Europe et l'Université du Luxembourg. En tant que porteur principal du projet, j'étais chargé d'en rédiger les conclusions. Au fait des pratiques de Twitter, nous avions fixé un *hashtag*, #dhlu2012. Le premier jour, nous pouvions suivre la conférence sur Twitter, sans que les tweets ne se distancient vraiment du contenu des interventions des participants. Le second jour était différent : nous avions affaire à deux conférences, l'une en réel, l'autre sur Twitter, rendant ce colloque ubiquitaire (Clavert, 2013). Les conversations sur le réseau partaient des contributions « réelles » et s'orientaient ensuite vers des thématiques connexes, mais non traitées pendant la conférence elle-même. Nous avions prévu de collecter les tweets, ce qui a permis d'en intégrer le contenu dans mes conclusions.

La même année, pour des raisons personnelles, j'ai essayé de collecter les tweets liés au hashtag #ledebat – 600 000 tweets émis pendant les trois heures du traditionnel débat entre les deux candidats à la présidentielle française admis au second tour. Les outils de collecte utilisés jusqu'alors adaptés aux conférences ne l'étaient pas pour une telle masse de données. Me renseignant plus avant sur les différentes manières de collecter des données, j'ai pu mettre au point un serveur usant cette fois du dispositif technique adéquat : j'ai ainsi pu collecter, toujours à titre personnel, plusieurs millions de tweets autour des hashtags #mariagepourtous et #manifpourtous lors du débat, en France, sur la réforme du mariage[11]. Avec cette collecte, j'ai tout simplement appris à utiliser l'API – je reviendrai sur ce terme – dite de *streaming* de Twitter. Le projet #ww1 était prêt à démarrer.

Le Centenaire et #ww1

Lancées en France le 11 novembre 2013, après quelques tribulations (Wieder, 2013), les commémorations pour le Centenaire de la Première Guerre mondiale voient,

---

[8] Les enregistrements vidéos de la conférence sont disponibles à l'adresse http://www.cvce.eu/recherche/unit-content/-/unit/c2586238-9c6c-4623-aa29-fe61e9b1a459.

[9] http://luxembourg2012.thatcamp.org/.

[10] http://tcp.hypotheses.org/category/thatcamp-paris-2012 — sont aussi consultables les traces des THATCamps Paris de 2010 et 2015. Les « non-actes » de THATCamp Paris 2012 sont publiés en ligne (*THATCamp Paris 2012*, 2012).

[11] Une analyse a été réalisée de ces tweets autour de la réforme du mariage, voir Cervulle & Pailler, 2014.

notamment en France, au Royaume-Uni et en Belgique[12], se développer de nombreux événements de toute nature autour du centenaire du déclenchement et du déroulement de la Grande Guerre. S'il a souvent été dit, avec raison, qu'en comparaison des dernières grandes commémorations de la Première Guerre mondiale, le Centenaire se déroulait dans un contexte nouveau, car plus aucun Ancien Combattant n'était encore en vie, une autre nouveauté est souvent oubliée : le contexte médiatique. En effet, les commémorations doivent désormais s'accommoder de l'émergence d'Internet, du web et des réseaux sociaux numériques comme Facebook, Twitter ou encore Pinterest, mais également des nombreux sites qui ont une « couche sociale » comme YouTube.

La collecte, par le biais de mots-clés, de tweets évoquant la Première Guerre mondiale ou son Centenaire n'a commencé que le 1er avril 2014, bénéficiant du savoir-faire acquis précédemment lors des quelques expériences mentionnées plus haut. Le grand intérêt scientifique d'une telle collecte est d'analyser la relation entre mémoire et histoire telle qu'elle s'exprime à l'ère numérique et de se demander si la perception du passé et sa mémoire évoluent avec les réseaux sociaux en ligne.

Sources primaires en flux

En d'autres termes, les aperçus introductifs ici développés nous poussent à nous poser une question fondamentale : comment fonder une recherche sur des sources primaires en flux ? Comment concilier la nature profonde de Twitter, c'est-à-dire un flot constant d'informations, avec ce que doit être un centre d'archives, c'est-à-dire un état relativement figé de traces du passé ?

2. Collecter

La collecte qui nous sert d'exemple ici a commencé le 1er avril 2014. Elle est multilingue, mais contient essentiellement des tweets en anglais et en français dans une moindre mesure. Commencée sur un mode artisanal, elle s'est professionnalisée depuis. Le script serveur utilisé collecte les données au format JSON, un format texte qui permet de structurer les informations, puis les stocke dans une base de données

[12] Pour une analyse du cas belge et une comparaison avec les autres nations, voir Bost & Kesteloot, 2014.

de type MySQL. La collecte est effectuée par le biais de mots-clés[13] : si un tweeta contient l'un des mots-clés collectés, il est alors automatiquement stocké dans la base.

Entre le 1er avril 2014 et le 13 avril 2016, 2 096 968 tweets ont été collectés. Le bruit, c'est-à-dire les tweets collectés mais n'évoquant pas la Grande Guerre ou son Centenaire, a été plutôt faible, à l'exception de « 11Nov » fortement utilisé par des indépendantistes catalans. Il est en augmentation depuis l'inclusion dans les mots-clés collectés de « Somme » et « Verdun », qui concernent autant le Centenaire que des actualités départementales et communales. Si l'on enlève les retweets, c'est-à-dire les citations telles quelles de tweets d'autres utilisateurs, cette base contient 730 111 tweets originaux.

Ces deux millions de tweets ont été émis par 542 570 comptes Twitter. Ces comptes sont de toutes sortes, dans la mesure où ils ont pu être ouverts par des individus, des institutions mémorielles, des médias, des projets de recherche. Certains peuvent être automatisés, comme @RealTimeWW1 qui est issu d'un projet pédagogique de l'Université du Luxembourg et émet tous les jours des tweets en piochant dans une base de données remplie par des étudiants.

Dans ces deux millions de tweets, on trouve 124 424 hashtags. Parmi ces nombreux hashtags, 54 566 n'apparaissent qu'une seule fois dans la base de données et 107 047 dix fois ou moins.

Problèmes méthodologiques liés à la collecte

La collecte de ces deux millions de tweets, qui continue toujours, n'est pas allée sans certains écueils.

Le premier est « classique » pour toute recherche reposant sur la constitution d'un corpus. Ce dernier n'existe que parce que cette recherche est menée et donne au corpus une certaine unité. Dans notre cas, par exemple, les visualisations réseaux (voir plus bas) rendent l'image d'une communauté qui s'est développée autour d'un même

---

[13] Les mots clés collectés sont les suivants : ww1, wwi, wwiafrica, 1gm, 1GM, 1wk, wk1, 1Weltkrieg, centenaire, centenaire14, centenaire1914, GrandeGuerre, centenaire2014, centenary, fww, WW1centenary, 1418Centenary, 1ereGuerreMondiale, WWIcentenary, 1j1p, 11NOV, 11novembre, WWI, poppies, WomenHeroesofWWI, womenofworldwarone, womenofww1, womenofwwi, womenww1, ww1athome, greatwar, 100years, firstworldwar, Verdun, Verdun2016, Somme, PoilusVerdun, Somme100. Certains de ces mots clés ont été collectés dès le départ, d'autres ont été intégrés plus tard, en fonction de l'évolution des commémorations, comme Somme100, utilisé pour le centenaire de la bataille de la Somme.

sujet. Cette communauté, toutefois, n'existe que parce que notre recherche lui donne une unité. Il faut ainsi éviter de surinterpréter les liens qui se nouent sur Twitter.

La question de la masse des données, ainsi que de l'éventuelle intégralité des données collectées, plonge l'historien dans les affres du *Big data*. Le *Big data* est une notion relative, dont on peut aisément trouver de multiples définitions, (par exemple dans Mayer-Schönberger & Cukier, 2013), et de nombreuses critiques (Boyd & Crawford, 2012). À l'origine défini selon des critères techniques – le fait de mettre au point les outils permettant d'exploiter en temps réel les incroyables masses de données engendrées par les grandes plateformes du Web comme les moteurs de recherche ou les réseaux sociaux numériques –, le *Big data* est désormais souvent utilisé dans les cas où il devient nécessaire de travailler avec un volume de données ne permettant plus une analyse strictement humaine mais forçant l'usage de l'informatique.

Ainsi, dans le cas de notre projet de recherche, nous ne pouvons strictement dire que nous travaillons avec des données massives de type *Big data*, il est néanmoins clair que nous devons faire face à des problématiques devenues aiguës avec le *Big data*. La première de ces problématiques est celle de l'« ordre illusoire ». Dans un article s'intéressant à l'usage par les historiens canadiens des journaux numérisés du XIX^e^ siècle, Ian Milligan remarque que les journaux francophones et locaux, peu numérisés mais bien sûr consultables en bibliothèque ou en centre d'archives, sont évincés au bénéfice des grands journaux, numérisés, et reflétant une vision canadienne centrale et anglophone du pays (Milligan, 2013). Il parle ainsi d'ordre illusoire : la consultation des grandes bases de données nous donne l'impression d'avoir consulté l'intégralité de nos sources alors qu'elle nous en fait oublier de nombreuses, rendues invisibles parce que non-numérisées et non-accessibles en ligne.

Dans le cas de notre recherche, travailler sur plusieurs millions de tweets à terme[14] ne doit pas nous faire oublier les angles morts de notre collecte. Nous avons obtenu des tweets sur la base de mots-clés. Cela signifie que de nombreuses personnes évoquant la Grande Guerre avec d'autres mots ne sont pas incluses dans cette base de données. La proportion de ces tweets non collectés est d'ailleurs difficile à estimer. À cela, nous devons ajouter que les analyses reposant sur des données issues de Twitter ne peuvent être extrapolées sans grandes difficultés pour comprendre d'autres réseaux sociaux

[14] La base de données rassemblait à la fin du mois de juillet 2016 environ 2,9 millions de tweets. Les éléments discutés dans la partie exposant quelques résultats sont issus d'une version de la base de données allant d'avril 2014 à avril 2016 et contenant environ deux millions de tweets. En effet, les mois de mai à juillet 2016, couvrant des commémorations importantes autour des batailles de Verdun et de la Somme, n'ont pas encore été analysés.

numériques[15], d'autres dispositifs socio-techniques du web comme les blogs ou les forums[16] ou encore ce qui se passe dans la « vie réelle ». Enfin, *Twitter Inc.* est une entreprise californienne, qui s'est ensuite développée dans le reste du monde : les taux d'usage par pays diffèrent fortement et notre pratique de la recherche doit en tenir compte. En somme, notre corpus ne peut être considéré comme exhaustif.

Nous devons également accorder une grande attention aux affres du flux. Pour les détailler, nous allons d'abord préciser une notion fondamentale, celle d'API. Une *Application Programming Interface* (API) ou interface de programmation est une partie d'un logiciel – ici, Twitter – qui permet d'échanger des fonctionnalités ou des données avec d'autres logiciels – dans notre cas, le script qui nous permet de collecter les données. En fonction de l'API choisie (Twitter en a plusieurs), les données récoltées ne sont pas les mêmes.

Dans le cas de Twitter, plusieurs API sont disponibles : l'API de recherche, qui permet de collecter jusqu'à 3000 tweets par heure émis dans les jours précédents la requête, l'API de *streaming* qui permet de collecter jusqu'à 1 % du total des tweets émis au moment de la requête en temps réel – on estime le nombre total de tweets émis à un demi-milliard par jour en 2015 (Blog du modérateur, 2015) – et la *full* API, accessible uniquement par la voie commerciale, qui permet de collecter tous les tweets, émis ou à émettre, nécessaires à une recherche. Il reste néanmoins que, pour de grandes collectes, l'usage de l'API *Streaming* ou même de la *full* API est soumise à un certain flou du côté de Twitter : même dans un cas où l'on est sous le 1 %, il n'est pas certain que l'on collecte l'intégralité des tweets contenant un mot-clé.

Twitter met en scène un flux constant d'informations. Tout notre projet de recherche repose sur l'idée de figer ce flux, en capturant des tweets dans une base de données.

---

[15] Ainsi, nos analyses montrent que les musées français utilisent mal Twitter, notamment car ils ne se plient pas à ses usages, comme l'utilisation d'un hashtag. Par exemple, le musée des Armées aux Invalides (@MuseeArmee) tweete la plupart du temps sans insérer de mots-dièse dans ses tweets. Cela revient en grande partie à communiquer dans le vide. Néanmoins, sans pouvoir pour le moment le prouver, il nous semble par exemple que les musées français utilisent bien mieux Facebook que Twitter pour leur communication. Ainsi le Musée de la Grande Guerre de Meaux a-t-il obtenu un très grand succès grâce au compte Facebook du Poilu fictif Léon Vivien (https://www.facebook.com/leon1914). Une chose ici doit être précisée. Nous travaillons sur des données issues de Twitter et non de Facebook, par exemple, car la possibilité nous en est offerte grâce à des API relativement ouvertes. Le projet Algopol (http://app.algopol.fr/info) collectait des données sur Facebook, jusqu'à un changement de la politique de ce réseau, qui a forcé ce projet de recherche à interrompre sa collecte plus tôt que prévu.

[16] Un autre projet de recherche, mis en œuvre par le LabEx *Les passés dans le présent* et par Télécom ParisTech, s'intéresse à un forum de passionnés de la Grande Guerre. Les échanges que nous avons pu avoir montrent que les utilisateurs des forums sur la Grande Guerre ne sont pas les mêmes que les utilisateurs de Twitter qui y parlent du premier conflit mondial.

Pour cela, nous devons utiliser les APIs de Twitter. La masse des données que nous collectons ne nous autorise pas à utiliser l'API de recherche, la seule à nous permettre, sans frais, de fouiller dans l'historique des tweets. En conséquence, nous devons utiliser l'API *Streaming*, c'est-à-dire que nous sommes forcés à anticiper les hashtags ou mot-clés, base de notre collecte. Ainsi, parce que les comptes Twitter francophones usaient de hashtags très précis, la première bataille de la Marne (septembre 1914) est-elle très peu présente dans notre base de données, bien que fondamentale, puisqu'elle marque l'arrêt de l'offensive allemande. Un autre exemple est la bataille du Bois Delville, qui a commencé le 14 juillet 1916 et est particulièrement importante pour les troupes sud-africaines. Pour sa commémoration, alors que cette bataille se joue dans le cadre de celle de la Somme, les hashtags utilisés ont été distincts et n'ont pas été spécifiquement collectés, ce qui risque d'amplifier encore l'importance des commémorations du 1er juillet 2016, centenaire du début de l'offensive de la Somme, aux dépens de commémorations plus modestes. Dans un tel projet de recherche, l'anticipation devient reine. Elle n'est toutefois pas toujours possible. Le délai entre la découverte d'un mot-dièse intéressant pour une recherche et sa prise en compte dans la collecte peut être très variable. S'il est très court et si le nombre de tweets est peu important, il est toujours possible de rattraper le « trou » créé par ce délai. Pour éviter une perte de données trop importante, il est ainsi nécessaire d'effectuer une veille constante.

Nous devons aussi considérer le choix du logiciel qui accède à l'API et rapatrie les données. L'outil que nous avons choisi pour la collecte, *140dev*, ne collecte pas les avatars des comptes Twitter : nous avons considéré que c'était en dehors du cadre de notre recherche. D'autres chercheurs n'auraient peut-être pas fait ce choix.

Enfin, cette recherche doit faire avec des données qui, aux yeux des historiens, sont finalement assez peu structurées (voir Figure 1). Nous sommes conscients que l'API de Twitter renvoie de nombreuses métadonnées : on peut en effet trouver des données bien moins structurées, particulièrement sur le web. Toutefois, prenons l'exemple des métadonnées à caractère temporel qui sont présentes dans notre base de données. Il y en a trois : la date de publication des tweets, la date de création des comptes Twitter et la date de dernière mise à jour de ces comptes.

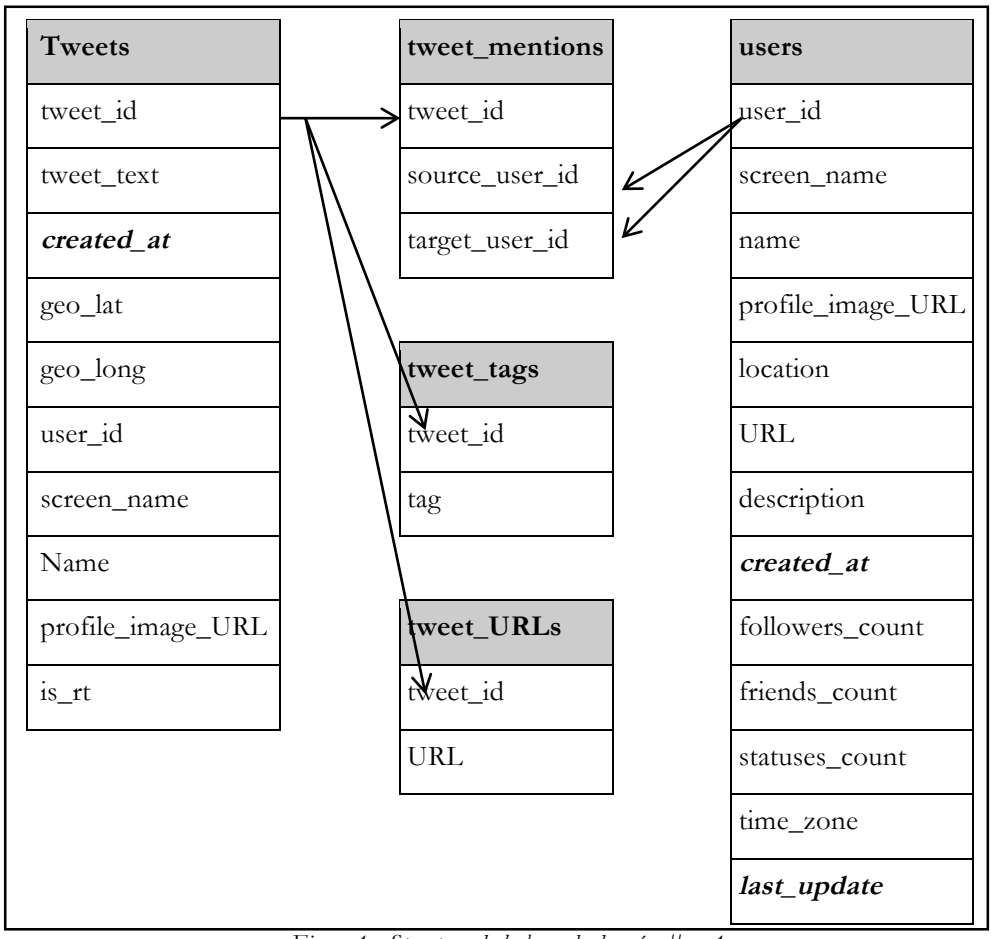

Figure 1 - Structure de la base de données #ww1
Les flèches représentent les liens d'une table de la base de données à une autre.
Les éléments en italique sont les champs contenant des informations temporelles.

Pourtant, on trouve, dans les tweets, dans les descriptions des comptes, bien plus de traces de temporalités que ces quelques métadonnées temporelles fournies par la plateforme. Dans le texte des tweets, des références de natures très diverses au temps (des dates très précises aux expressions temporelles vagues comme « jadis ») sont présentes, reflétant des temporalités qui s'imbriquent les unes avec les autres : les temporalités de la Grande Guerre, du Centenaire, de chaque utilisateur, du flux qu'est Twitter. Les métadonnées fournies par Twitter reflètent une vision linéaire et non finie du temps qui n'est pas nécessairement celle qui est exprimée dans les tweets.

Comprendre ces références temporelles nécessite de faire appel à des techniques informatiques dites de reconnaissance d'entités nommées dont la qualité est aujourd'hui très variable bien qu'en progrès constants[17].

Ces embûches méthodologiques liées à la collecte ne sont pas les seules. L'analyse des tweets une fois collectés en recèle quelques autres.

3. Analyser : l'historien.ne face à une mer de données

Une fois les tweets collectés, la grande question qui se pose pour leur analyse est la suivante : comment lire deux millions de tweets ? Que peut faire un.e historien.ne face à une mer de données ?

Nous recourons à une notion qui a été avancée par l'historien de la littérature Franco Moretti : la lecture distante ou *Distant Reading*. Moretti l'introduit dans son ouvrage *Graphs, Maps and Trees*, en s'interrogeant sur la manière de faire l'histoire de la littérature (Moretti, 2007) : faut-il faire l'histoire des grands romans ou essayer d'embrasser toute la littérature, y compris les textes oubliés ou considérés comme mineurs ? Moretti, souhaitant aller sur cette seconde piste, est confronté à une masse d'informations qu'il ne peut gérer et interpréter sans appel à une médiation computationnelle. Il propose de s'inspirer de l'histoire économique quantitative de l'École des Annales (*Graphs*), des géographes (*Maps*) et de la théorie de l'évolution (*Trees*) pour dégager les grands traits de grandes familles de littérature et comprendre les liens entre elles. Il s'agit ainsi, en utilisant des techniques informatiques et des méthodes provenant d'autres disciplines, de se distancier de ses sources pour en embrasser un regard plus englobant et pouvoir en analyser de plus grands ensembles.

Notre démarche a suivi cette logique, avec un bémol : à chaque étape de notre recherche, nous pouvons à la fois regarder l'ensemble des tweets collectés et retourner à des groupes de tweets plus précis, voire à un tweet en particulier. Le 11 novembre 2014, nous avons pu retracer la destinée d'un tweet singulier, qui citait une lettre de Poilus vraisemblablement fausse, bien que le tweet soit de toute bonne foi. Ainsi, avons-nous observé la recherche d'informations de nombreux membres de Twitter s'interrogeant sur l'authenticité de la lettre. Ce même jour, nous avons également analysé les grands axes du contenu de plus de 90 000 tweets.

---

[17] C'est une méthode que nous n'avons pour le moment pas encore entreprise.

Les technologies permettant d'exercer cette lecture distante ressortent, dans notre cas, essentiellement de la fouille de texte (*text mining*) et de l'analyse et visualisation des réseaux sociaux.

Quelques résultats

Nous nous intéressons en premier lieu au flux de tweets sur la Première Guerre mondiale, en regardant simplement le nombre de tweets par jour (Figure 2).

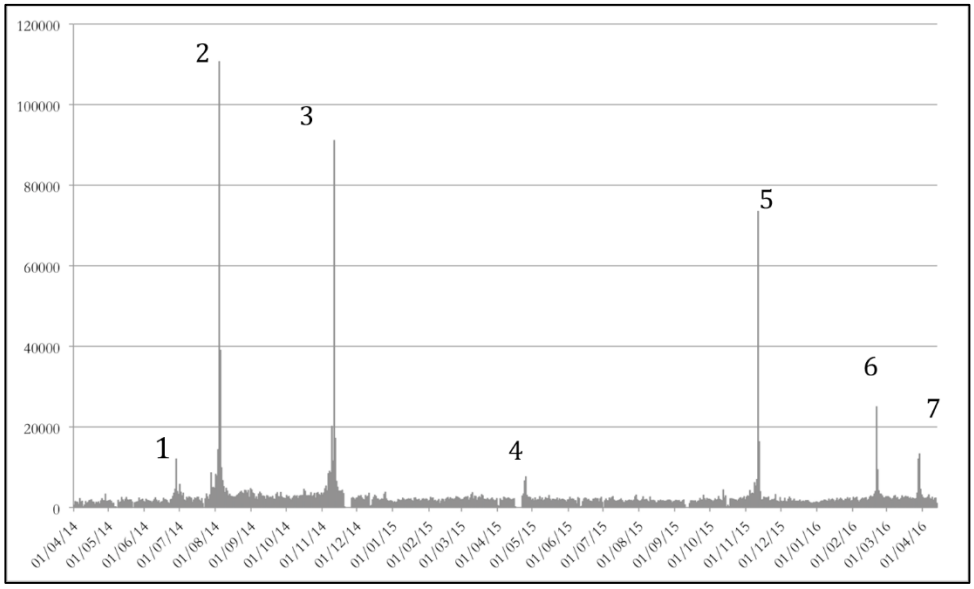

Figure 2 - Nombre de tweets par jour

Ce simple graphique nous permet de repérer les événements les plus marquants des échos du Centenaire sur Twitter :

1. Centenaire de l'assassinat de l'Archiduc François-Ferdinand, 28 juin 2014, 12 029 tweets ;
2. Centenaire de l'entrée en guerre du Royaume-Uni, 4 août 2014, 110 684 tweets ;
3. Commémoration de l'armistice, 11 novembre 2014, 91 108 tweets ;

4. ANZAC Day, 25 avril 2015, 7 605 tweets ;
5. Commémoration de l'armistice, 11 novembre 2015, 73 520 tweets ;
6. Centenaire du déclenchement de la guerre de Verdun, 21 février 2016, 25 011 tweets ;
7. *Easter rising*, 28 et 29 mars 2016, 12 021 et 13 299 tweets respectivement.

On peut ensuite tenter de comprendre la répartition linguistique de ce flux, ce qui nous permet de constater que, des sept pics repérés plus haut, trois sont francophones (les 11 novembre 2014 et 2015 et le centenaire de la bataille de Verdun). Seul le 11 novembre est un pic récurrent.

Une fois cette simple analyse du flux effectuée, nous procédons à une analyse de texte. Nous nous limiterons ici à l'analyse du corpus francophone (Figure 3). Pour procéder à cette analyse, nous utilisons un logiciel, IRaMuTeQ [18], qui implémente la classification hiérarchique descendante (Ratinaud & Dejean, 2009) qui avait été mise au point par Max Reinert (Reinert, 1993) et était déjà disponible avec le logiciel Alceste[19]. Pour décrire rapidement la méthode, il s'agit de regrouper dans des classes ou profils des tweets se ressemblant, sur la base de la concurrence de mots dans les tweets. Si les tweets dans une classe ont des similarités, la division des classes se fait sur la base de dissemblances d'une classe à l'autre. Plus clairement, cela signifie que chaque classe visualisée dans la figure qui suit correspond à un grand thème présent dans les tweets. Les mots qui sont présents sous les classes sont les mots les plus représentatifs de ces classes.

[18] http://www.iramuteq.org/. IRaMuTeQ est un logiciel libre.

[19] http://www.image-zafar.com/. Alceste est un logiciel commercial.

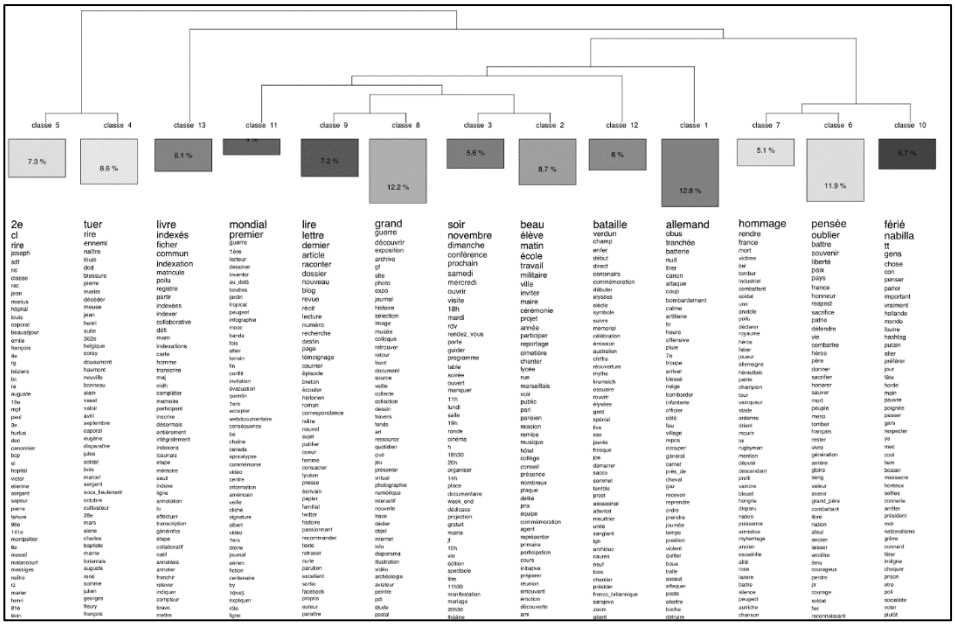

Figure 3 - Classification hiérarchique descendante (méthode Reinert) du corpus francophone (sans les retweets)

Ce que montre une telle analyse, ce sont les grands thèmes abordés dans les tweets francophones sur le Centenaire. Par rapport au corpus anglophone, par exemple, les tweets francophones se préoccupent nettement plus des Poilus. Les anglophones semblent plus tournés vers la mémoire du champ de bataille. L'entrée en guerre (4 août 1914) a fait l'objet, également, de nettement plus d'attention que la sortie de guerre (11 novembre 1918) dans le corpus anglophone alors que le corpus francophone est plus tourné vers les célébrations du 11 novembre. Absente sur ce graphique et non encore analysée, la commémoration du centenaire de la bataille de la Somme, organisée par les autorités françaises et britanniques, le 1er juillet 2016 sera peut-être un moment commun des aires linguistiques. C'est par ailleurs, pour le moment, le jour où la Grande Guerre a été la plus twittée.

Comme nous avons gardé, dans notre analyse, les dates d'émission des tweets, nous pouvons projeter dans le temps ces grands thèmes révélés par l'analyse de texte. Nous constatons alors que les tweets francophones évoquent toute l'année les Poilus, mais avec des mots différents en fonction de la date. Ainsi, le 11 novembre et les jours de grande commémoration sont-ils marqués par un hommage général aux Poilus morts pour la France. Le reste de l'année, tout en restant dans une forme d'hommage, ce sont les parcours individuels des soldats tombés sur le champ de bataille qui sont contenus dans les tweets. On peut d'ailleurs remarquer l'activisme mémoriel qui se

déploie autour de la base de données des morts pour la France, fondamentale pour comprendre l'activité des comptes Twitter francophones[20].

Outre l'analyse de contenu, il est intéressant de comprendre également qui twitte et quels sont les comptes Twitter les plus actifs. Pour cela, on peut regarder la visualisation réseau formée par les mentions et retweets (Figure 4). Ce qui frappe est la centralisation côté français autour de la mission du Centenaire (*@mission1418*), qui coordonne les manifestations du Centenaire dans le groupe francophone (1) et la diversité des comptes côté anglais (2) : des médias avec la BBC, des musées avec les Imperial War Museums (*@I_W_M*), des organisations d'anciens combattants (*@PoppyLegion*), des projets d'histoire publique (*@letter1418*), des amateurs d'histoire (*@HistoryNeedsYou*). On peut toutefois, côté francophone, remarquer le compte *@1j1Poilu* qui motive l'ensemble des personnes en train d'indexer la base de données des Morts pour la France.

[20] La base des données des Morts pour la France est éditée par le Ministère français de la Défense. Elle contient les reproductions des actes administratifs déclarant un soldat « mort pour la France ». À l'origine très pauvre en métadonnées, elle propose un module permettant à tout un chacun d'indexer la base, c'est-à-dire de transformer une image des actes administratifs en texte et métadonnées.

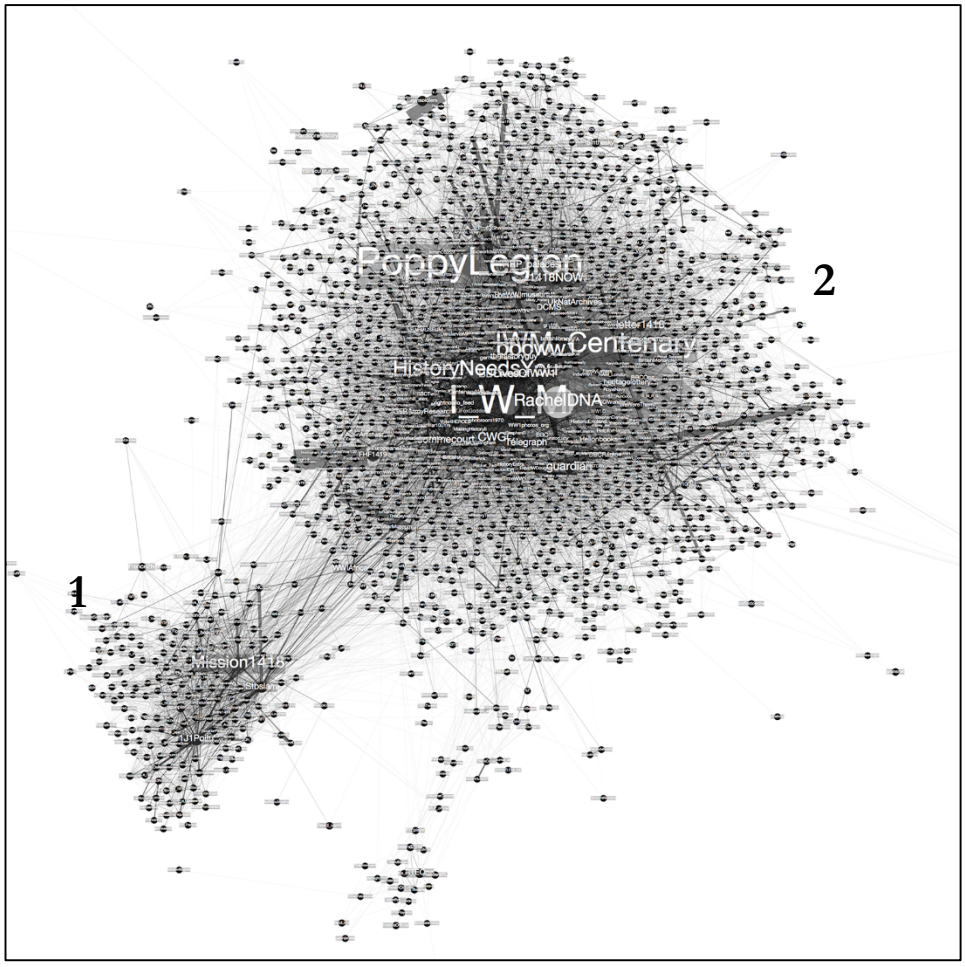

Figure 4 - Graphe des mentions et retweets
(limité aux comptes mentionnés / ayant mentionnés ou retweetés / ayant retweetés au moins 15 fois)

Limites de cette approche

Les quelques résultats que nous avons rapidement exposés ici nous renvoient aux éléments méthodologiques décrits plus haut, ce qui nous permet de mieux appréhender les limites de notre approche.

Le premier ensemble de limites est technique. Comme nous l'avons déjà noté plus haut, les données sont, au regard de l'historien, finalement assez peu structurées. Nous devons également prendre en compte les limitations liées aux conditions de la

collecte : une telle recherche nécessite des moyens matériels qui ne sont pas toujours à disposition des chercheurs. Dans notre cas, la collecte a commencé de manière artisanale, par l'usage d'un serveur d'occasion en auto-hébergement dépendant de la qualité fort relative de notre connexion Internet privée. Depuis avril 2016, le serveur a été migré vers une machine virtuelle hébergée à l'Université de Lausanne, qui dispose d'une infrastructure technique fiable, ce qui comprend la sauvegarde des données collectées, mais ne comprend pas leur pérennisation au-delà de quelques années. Toutes les universités ne disposent pas de cette infrastructure et, quand elles en disposent, ne la proposent pas systématiquement gratuitement aux chercheurs qui en ont besoin.

Le second ensemble de limites est relatif à Twitter. Outre les remarques générales déjà effectuées plus haut, nous avons pu constater des interruptions dans la collecte de tweets, liées à la défaillance de l'API de Twitter ou à un changement technique de cette API. Ces interruptions ont pu durer jusqu'à quelques jours, mais, vérifications faites, ne nous ont pas fait manquer un événement particulièrement important. Notons aussi que la société *Twitter Inc.* rencontre des difficultés sérieuses : ce projet de recherche peut être interrompu à tout moment par sa faillite ou, plus vraisemblablement, par son rachat par une société qui changerait la politique d'accès aux données.

Enfin, certaines limites sont liées à l'historien.ne, à sa formation initiale et au temps qu'il.elle peut consacrer à l'acquisition de nouvelles compétences techniques, ou, en cas de collaboration avec des ingénieurs ou des chercheurs en sciences informatiques, d'un langage commun, c'est-à-dire d'une culture générale numérique. L'usage des logiciels que nous employons (tableur, Gephi[21], IRaMuTeQ ou OpenRefine[22] pour donner quelques exemples) implique un temps d'acquisition des savoir-faire qui est particulièrement long avec une courbe d'apprentissage qui peut être aiguë et, comme pour la collecte, l'accès à une infrastructure numérique qui est inégal d'une université à l'autre. Plus compliquée encore est l'acquisition des éléments méthodologiques et conceptuels issus de la statistique, de la sociologie ou des sciences de l'information. L'historien.ne est appelé, sur ce genre de projet, à accentuer ses efforts de coopération avec chercheurs, ingénieurs, bibliothécaires et archivistes issus d'horizons différents du sien.

Notons également que les outils d'analyse peuvent être difficiles à choisir. Ces logiciels sont fondés sur des hypothèses scientifiques. Ainsi, deux logiciels d'analyse de texte

---

[21] Gephi nous sert à procéder aux analyses réseaux et à produire les visualisations. https://gephi.org/.

[22] OpenRefine nous permet de travailler nos données pour les rendre utilisables par les logiciels d'analyse. http://openrefine.org/.

ne produiront que très rarement des analyses réellement similaires. Le choix d'un logiciel correspond aussi au choix d'une école disciplinaire.

4. Archiver et pérenniser : que faire des données ?

Collecte et analyse étant effectuées, que faire pour archiver et pérenniser le projet et ses données et, ainsi, garder une certaine crédibilité des résultats issus des analyses ? La question de l'archivage est difficile dès qu'il s'agit du web.

On peut considérer plusieurs niveaux d'archivage : l'archivage du web, l'archivage des réseaux sociaux numériques dans leur ensemble, l'archivage des données liées à un projet de recherche ou à un.e chercheur.euse.

L'archivage du web, bien souvent, ne comprend pas, par défaut, l'archivage des réseaux sociaux numériques. *Internet Archive*, un organisme privé d'archivage du web, ne se préoccupe pas des réseaux sociaux numériques[23]. En France, la Bibliothèque nationale de France et l'Institut National de l'Audiovisuel, qui sont en charge du dépôt légal du web, couvrent en premier lieu le domaine « .fr » et une sélection de sites web hors de ce domaine. Ponctuellement, ils peuvent également archiver des tweets, par exemple à la suite des attentats de 2015 en France. Dans le cas suisse, l'archivage du web est assuré par la bibliothèque fédérale en coopération avec les bibliothèques cantonales. La politique d'archivage est sélective, ne comprend pas de collecte systématique du « .ch » et exclut les réseaux sociaux numériques[24].

Au niveau des réseaux sociaux numériques, les politiques d'archivage sont des plus floues. Facebook s'archive lui-même, par exemple. Le spectre de la disparition avec armes, bagages et archives flotte au-dessus de ces services : le cas récent de la fermeture (entendons, l'effacement) de l'ensemble des sites de *GeoCities* par leur propriétaire, *Yahoo, Inc.*, dont l'archive a été sauvée *in extremis* nous le rappelle (Archiveteam, 2009).

La grande exception est Twitter. Ce dernier est en effet archivé – même si certaines modalités de l'archivage restent inconnues – par la Bibliothèque du Congrès des États-Unis (Library of Congress, 2013)[25]. Les tweets publics sont ainsi conservés depuis 2006 par une institution publique. Pour le moment, toutefois, cette archive n'est pas consultable. Les difficultés techniques liées à une mise à disposition aux chercheurs sont trop importantes (Zimmer, 2015).

---

[23] Sur l'archivage du web, on consultera à sa parution (Brügger, 2017).

[24] D'après https://www.nb.admin.ch/nb_professionnel/01693/01695/01705/03333/index.html.

[25] Voir la contribution de J.-D. Zeller dans le présent ouvrage.

Le troisième type d'archivage se situe au niveau des chercheurs et projets de recherche. La question de l'archivage des données de la recherche est fondamentale, car en dépend la reproductibilité des résultats de la recherche et, en conséquence, la crédibilité de nombreuses publications scientifiques. Elle contient plusieurs dimensions : il s'agit à la fois de l'archivage des données et des logiciels pour les exploiter. Beaucoup de bases de données d'historiens élaborées dans les années 1990 sont aujourd'hui perdues, car elles ne sont pas lisibles. Cet archivage nécessite le recours à des infrastructures nationales, européennes ou internationales – l'un des meilleurs exemples étant celui de la Très grande infrastructure de recherche Huma-Num en France[26].

Toutefois, archiver des tweets collectés par des chercheurs pose des problèmes spécifiques. D'une part, les conditions d'utilisation de Twitter sont assez restrictives : les données collectées par des chercheurs (ou des entreprises) ne peuvent être mises à disposition du public. Leur archivage ne signifie ainsi pas qu'elles seront accessibles. D'autre part, l'archivage d'une base de données de tweets questionne sur ce qui doit être archivé. Dans le cas des utilisateurs, la question de l'autorité est déterminante (Merzeau, 2013) : un tweet n'a pas la même autorité selon le compte Twitter qui l'a émis et, notamment, en fonction du nombre de ses *followers* par exemple. Lorsque l'on veut estimer cette autorité, on peut regarder les statistiques du tweet fournies par Twitter. Ces statistiques peuvent-elles être archivées et dans quelles conditions ? Ces statistiques incluent notamment les « impressions », qui peuvent donner une idée du comportement des comptes qui ne publient pas de tweets et se contentant de lire ceux des autres. Un autre élément qui questionne l'archivage de Twitter – par le biais de la Bibliothèque du Congrès ou de l'archivage des données de la recherche – est l'archivage des interfaces. On peut accéder à Twitter avec différents logiciels, avec différents types de terminaux informatiques (ordinateur, tablette, téléphone). L'interface est constitutive de la perception d'un tweet par un utilisateur.

La problématique que soulève l'archivage de Twitter est cependant plus vaste encore : comment archiver tout un écosystème numérique ? Twitter, avec ses APIs plutôt ouvertes, a engendré la création de tout un environnement. De nombreux sites vous permettent d'user de Twitter de manières très différentes, influençant la manière dont vous rédigez vos tweets (par exemple en faisant recours à un racourcisseur d'URLs),

---

[26] Le site de la Très grande infrastructure de recherche Huma-Num est disponible à l'adresse http://www.huma-num.fr/. Les outils et services qu'elle fourni sont décrits ici : http://www.huma-num.fr/services-et-outils et sont centrés sur les données de la recherche, qu'il s'agit de « stocker », « traiter », « diffuser », « archiver », « signaler » et « exposer ». Parmi les importantes contributions d'Huma-Num, signalons un moteur de recherche centré sur les sciences humaines et sociales, Isidore : https://www.rechercheisidore.fr/.

et, bien sûr, le contenu des tweets. Twitter encourage, avec sa limitation à 140 caractères par tweets, le partage des liens. Ces derniers sont-ils systématiquement archivés par un service, public ou privé, d'archivage du web ? Rien n'est moins sûr : après un an, de nombreux liens partagés sur les réseaux sociaux pendant le Printemps arabe étaient perdus (SalahEldeen & Nelson, 2012).

5. Conclusion

Au cours de la recherche exposée ici, de nombreuses difficultés ont été rencontrées. Pendant la collecte, les affres du *Big data*, doivent être prises en compte : outre les difficultés techniques qui peuvent être liées à la gestion d'une grande masse de données, cette dernière ne doit pas nous faire croire à un ordre illusoire, c'est-à-dire l'impression d'avoir constitué un corpus tellement imposant qu'il engloberait nécessairement l'ensemble des données nous intéressant. Beaucoup d'éléments ne peuvent être intégrés au corpus et ceci doit nous inciter à une certaine humilité, seul rempart contre un *hubris* du *Big data*. Les affres du flux doivent aussi être prises en compte : l'anticipation dans la collecte des données – notamment dans le choix des hashtags -, est reine mais n'est pas toujours possible, menant à des absences dans le corpus qui ne sont pas toujours compensables. Enfin, la masse des données ne doit pas nous faire oublier que les métadonnées des tweets n'ont pas été pensées pour un usage historien. Elles sont loin, notamment, de refléter la complexité des temporalités s'exprimant dans les tweets.

Un second ensemble d'embûches réside dans l'exploitation scientifique des tweets collectés. Comment analyser une masse de données qui, non seulement interdit la lecture humaine exhaustive d'un corpus, mais en outre augmente régulièrement au fur et à mesure de la recherche ? Nous devons faire appel à la notion de lecture distante, c'est-à-dire introduire une médiation informatique dans notre analyse et, partant, dans l'interprétation des tweets. Travailler sur des données issues de Twitter met en lumière certaines limites de l'historien.ne et de sa formation, limites qui doivent nous pousser à revoir nos relations avec nos collègues archivistes, ingénieur.e.s et informaticien.ne.s, parfois plus à même de traiter ce type de données.

Enfin, un troisième groupe de difficultés émerge, touchant à l'archivage et à la pérennisation des tweets. Si Twitter est une exception, car archivé par une institution publique reconnue, ce service a engendré un écosystème, c'est-à-dire un ensemble de sites web, de pratiques, de logiciels que l'existence de ce réseau social numérique entretient. Il est extrêmement difficile, voire impossible, d'archiver l'ensemble de cet écosystème.

Pour partie, ces problématiques s'expliquent par une forme d'antagonisme, de tension entre un flux constant d'informations, Twitter, et une archive – dans le cas présent, notre base de données. Le flux redonne à la métaphore de « source » un sens plus originel. Vouloir collecter ce flux pour ses recherches, l'analyser puis l'archiver revient à vouloir le figer, engendrant les nombreuses difficultés, techniques, méthodologiques que nous rencontrons au quotidien.

Ces difficultés méthodologiques et techniques ne doivent toutefois pas dissuader les historien.ne.s et archivistes de travailler sur de tels types de source. Dans le cas du projet de recherche exposé dans ce chapitre, les enjeux de l'ère du flux ne peuvent pas laisser indifférent, tant Twitter et les réseaux sociaux numériques sont susceptibles de modifier les temporalités de nos mémoires collectives.

Bibliographie

Archiveteam. (2009). *GeoCities*. Consulté 19 juillet 2016, à l'adresse : http://www.archiveteam.org/index.php?title=GeoCities

Blog du modérateur. (2015). *Chiffres Twitter - 2015*. Consulté à l'adresse : http://www.blogdumoderateur.com/chiffres-twitter/.

Bost, M., & Kesteloot, C. (2014). Les commémorations du centenaire de la Première Guerre mondiale. *Courrier hebdomadaire du CRISP*, *30-31*(2235-2236), 5-63.

Boullier, D. (2015). Les sciences sociales face aux traces du Big data. *Revue française de science politique*, *65*(5), 805–828.

Boyd, D., & Crawford, K. (2012). Critical uestions for Big data : Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679. http://doi.org/10.1080/1369118X.2012.678878.

Brügger, N. (2017). *The Archived Web: Doing History in the Digital Age*. MIT Press.

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). Consulté à l'adresse : http://rfsic.revues.org/717

Clavert, F. (2013). *Blogs et réseaux sociaux : des conférences scientifiques ubiquitaires ?* Consulté 19 juillet 2016, à l'adresse : https://2013.geschichtstage.ch/referat/368/blogs-et-rseaux-sociaux--des-confrences-scientifiques-ubiquitaires-

Library of Congress. (2013). *Update : Twitter Archives, Library of Congress* (p. 5). Washington: Library of Congress. Consulté à l'adresse : http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Boston : Houghton Mifflin Harcourt.

Merzeau, L. (2013). Twitter, machine à faire et défaire l'autorité. *Médium*, *1*(34), 171-185.

Milligan, I. (2013). Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010. *Canadian Historical Review*, *94*(4), 540-569. http://doi.org/10.3138/chr.694

Moretti, F. (2007). *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso.

Ratinaud, P., & Dejean, S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. In *Modélisation Appliquée aux Sciences Humaines et Sociales*. Toulouse. Consulté à l'adresse http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf/view

Reinert, M. (1993). Les « mondes lexicaux » et leur « logique » à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, *66*(1), 5-39. http://doi.org/10.3406/lsoc.1993.2632

SalahEldeen, H. M., & Nelson, M. L. (2012). Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? *arXiv:1209.3026*. Consulté à l'adresse http://arxiv.org/abs/1209.3026

THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques. (2012). Éditions de la Maison des sciences de l'homme. Consulté à l'adresse http://books.openedition.org/editionsmsh/278

Wieder, T. (2013). Généalogie heurtée d'un « événement majeur ». *Le Débat*, *176*(4), 160. http://doi.org/10.3917/deba.176.0160

Zimmer, M. (2015). The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*, *20*(7). Consulté à l'adresse http://firstmonday.org/ojs/index.php/fm/article/view/5619