Au-delà des traces numériques visibles

Mariannig Le Béchec, Camille Alloing

Au-delà des traces numériques visibles

Mariannig LE BÉCHEC
Camille ALLOING

Université de Poitiers/CEREGE (EA 1722)

France

Référence : Le Béchec Mariannig, Alloing Camille, « Au-delà de straces numériques visibles », in Zlitni Sami, Liénard Fabien, *Réseaux sociaux, traces numériques et cmmunication électronique*, Le Havre, 2018

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

Introduction

Lorsqu'une entreprise offre à des chercheurs une copie de sa base de données, c'est une sorte de graal qui semble leur tomber dans les mains. Nous allons enfin pouvoir observer ce qui se passe au-delà de l'interface que nous pouvons consulter page par page à la recherche de traces d'usage. Voire, les chercheurs imaginent que ces données pourront être aisément croisées avec d'autres données extraites à partir d'*Application Programming Interface* (API) de plateformes telles que Twitter, Facebook et YouTube. Notre hypothèse est d'observer à partir de ces méthodes la localisation du financement participatif en France afin d'en comprendre ses petits mondes (Le Béchec *et al.*, 2017). Au final, l'observation et l'analyse de ces traces se sont avérées beaucoup plus complexes à mettre en œuvre, quand bien même, les jeux de données et l'extraction de données avaient été préalablement vérifiés. Ainsi nous proposons de nous interroger sur ces méthodes d'analyse des traces autour de la problématique suivante. Pourquoi les chercheurs étudiant une participation collective sur des plateformes doivent-ils interroger leurs méthodes d'analyse des traces? Nous

montrerons dans ce chapitre que loin d'être visibles sur les interfaces, les traces sont produites par des infrastructures, au sens des *infrastructures studies,* et induisent des transpositions dans nos méthodes et analyses.

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

Quelle ingénierie de projet pour analyser des traces ?

Il serait aisé de croire que tout peut être tracé et que nous pouvons récupérer ces traces, ou leur pendant quantitatif, des données. Mais l'usage des API nous montre que l'extraction de données sur des activités passées est limitée, que l'accès à des données ouvertes comporte plus que des lacunes et que le traitement de ces données tant du point de vue de la déclaration que de la confidentialité est une question encore émergente en sciences de l'information et de la communication.

Peut-on extraire des données via les API des plateformes pour répondre aux hypothèses ?

Si Marres et Gerlitz (2016) définissent les méthodes d'interfaces comme cette incroyable concordance des objectifs et des pratiques de recherche avec le design des interfaces, nous nous interrogeons ici sur la construction des hypothèses de recherche via des API. Ces interfaces, véritables « vues utilisateur » (Dagiral, Peerbaye, 2013) sont mises à disposition des développeurs, par les plateformes à la recherche d'innovation (Helmond, 2015). Nous les utilisons ici à des fins de recherche, c'est-à-dire dans le but d'obtenir des indicateurs d'analyse cohérents, que l'on peut répliquer d'une plateforme à une autre. Notre objectif était donc de trouver des localisations concordantes entre les réseaux sociaux des porteurs de projets de financement participatif sur les plateformes et les participants à leur campagne de financement

participatif. Un premier travail a donc consisté à consulter les API des plateformes Facebook[1], YouTube[2] et Twitter[3] pour vérifier les extractions de données possibles. Un second travail à consister à estimer le nombre de messages concernés (posts et commentaires, vidéos et commentaires, tweets et retweets, commentaires) et à relever dans un tableur les URL des posts, vidéos ou tweets.

La question devient, comme dans tout management de projet, contrainte par deux variables: le temps et le coût. Pouvons-nous le répliquer et devons-nous recourir à un développeur ou est-ce soutenable manuellement ? Pour deux plateformes, le travail du développeur a consisté à construire un référentiel de données puis à élaborer un algorithme d'extraction et d'insertion des données des plateformes dans ce référentiel. Les partenaires du projet ont fourni une copie de leurs bases de données anonymisées sous forme de fichiers texte. Chaque jeu de données a été réintégré dans un schéma de base de données relationnelle. Une deuxième base de données a été initialisée avec le référentiel des communes de France qui a dû être vérifié manuellement du fait de regroupement de communes. Enfin, une troisième et dernière base de données a servi à exploiter les données extraites via les API. Parmi l'ensemble des projets financés sur les plateformes des partenaires, nous avons retenu cinq projets culturels. L'idée de base était d'extraire des réseaux sociaux l'ensemble des informations exposées publiquement pour chacun de ces projets. Pour questionner Facebook, on utilise une application (pas en production) et son *token* ou jeton d'identification associé pour interroger l'API Graph. À chaque type de données renvoyé par l'API correspond une fonction de requête qui permet d'obtenir des informations pour ce type et une fonction de *mapping* qui le transforme dans un objet issu de notre modèle de données.

[1] https://developers.facebook.com/docs/graph-api

[2] https://developers.google.com/apis-explorer/?hl=fr#p/youtube/v3/

[3] https://developer.twitter.com/en/docs/basics/rate-limiting

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

Si Facebook et YouTube autorisent ces extractions, Twitter ne permet pas de remonter plus de 5 jours dans le passé. La difficulté d'observation des « traces-artefacts » est de vouloir observer la présence du passé (Jeanneret, 2011, p. 62). Ainsi, nous présenterons ultérieurement comment nous avons braconné pour obtenir ces données.

Comment peut-on utiliser ces traces ?

Outre les restrictions d'usage d'une copie de base de données commerciale, des accès aux API, les chercheurs doivent pour la mise en œuvre de la recherche se conformer aux contraintes juridiques de propriété intellectuelle, de protection des données personnelles et techniques de leur propre structure d'accueil.

Une fois les premiers jeux de données obtenus, les bases de données commerciales et institutionnelles, qualifiées parfois de *hard data*, doivent être agrégées dans une architecture de données pour permettre leur enrichissement avec des données sociales extraites du web et qui sont qualifiées de *small data* (boyd, Crawford, 2012). Tout un jeu de traduction s'opère alors entre les chercheurs et les prestataires en informatique, issus du monde industriel. Pour parvenir à la création d'un site web de stockage et de consultation des données accessible à tous les chercheurs, le responsable de projet doit alors :

- démarcher le directeur des systèmes d'information (DSI) de l'université pour obtenir une infrastructure matérielle et logicielle ;

- obtenir une validation de la responsable Commission nationale de l'informatique et des libertés (CNIL) de l'université pour l'anonymisation et la destruction des données et remplir une déclaration de traitement de données ;

- établir des contrats de confidentialité avec les enquêtés pour développer une observation ethnographique à partir de leur propre compte sur les trois plateformes ;

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

- négocier la validité de ces contrats liés aux enquêtés avec la cellule valorisation de son université qui n'a jamais vu ces types de contrat ;

- négocier des contrats de confidentialité des prestataires extérieurs selon les règles de la propriété intellectuelle et des bases de données avec la cellule valorisation de son université.

À ce stade, le chercheur qui travaille sur les données commerciales s'aperçoit que les données comportent des erreurs dues à l'extraction faite par l'entreprise et également, dues à la malice des internautes clients indiquant comme commune d'habitation « compote de fraise », obligeant à des traitements manuels et algorithmiques pour corriger ces données. La collecte, le nettoyage et l'analyse des données sont des méthodes éprouvées en fouille de données. La réunion de différents jeux de données oblige à homogénéiser ces données par rapport à son propre référentiel de données.

Nous sommes revenus à des méthodes éprouvées de traitement et de représentation des résultats de recherche tout en abandonnant certaines hypothèses et méthodes d'analyse. Les manières de faire sont donc nouvelles avec le recours à des outils informatiques pour extraire des données mais pour restituer ces résultats et traiter leur validité, nous sommes revenues à des méthodes qualitatives éprouvées.

Pourtant comme nous avons déjà pu le souligner (Le Béchec, Alloing, 2016), les attributs pour définir un territoire ou une organisation dans ce type de corpus dépassent les catégories communément admises de description du monde social.

Ce que les traces et les données ne disent pas

Notre constat est que les plateformes nous empêchent de stabiliser nos méthodes d'extraction des données. Il convient alors de développer des tactiques, usée de *métis*

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

au sens de Michel de Certeau (1990) afin de pouvoir extraire des données. Ces tactiques deviennent de véritables bricolages méthodologiques.

L'analyse multiplateforme des traces

L'analyse multiplateforme a porté également sur Twitter. Depuis 2012, cette plateforme a réduit considérablement l'accès à ses API, amenant à penser de nouvelles manière de consommer l'information (Alloing, 2014). Cette limitation nous amène à deux solutions possibles. La première consiste à *scapper,* littéralement aspirer, les messages (tweets) et leurs contenus (images, vidéos, etc.) ou métadonnées (dates, heures, localisation, etc.) en direct avec un historique de 5 jours maximum. Pour la seconde, il faut recourir à un prestataire agréé, un *data brocker* comme Datasift[4], ou utiliser une prestation proposée par Twitter[5] afin d'accéder de manière rétroactive à des corpus pré-définis. Ces tactiques permettent de souligner un premier travail invisible sur les données qui influe sur nos analyses. Ce travail permet de transformer des données « métiers », utiles aux *data brockers* et aux ingénieurs de Twitter, en données « brutes » qui nous serviront ensuite. Cette « brutification » sert à dé-corréler les usages initiaux des données de leurs futurs usages, leur accordant ainsi une certaine agentivité. *« L'enjeu des transformations [que les données] subissent n'est plus de corriger des biais de mesure ou de distinguer le bruit de l'information à traiter (comme en sciences), mais d'assurer une migration d'une donnée « étroite », locale, vers une donnée à vocation universelle. »* (Denis, Goëta, 2013, p.16).

Les messages pertinents pour nos travaux datant de plus de 4 ans et notre budget étant limité, aucune des deux premières solutions n'a été retenue. Nous avons donc développé notre propre tactique (de Certeau, 1990) à partir du moteur de recherche de

[4] https://datasift.com/

[5] https://developer.twitter.com/en/enterprise

Twitter pour accéder aux messages voulus. Cet outil propose des résultats éditoriaux: mise en visibilité en fonction de critères propres à l'usager faisant la requête (localisation, graphe relationnel, historique, ciblage publicitaire) et agencement en fonction de critères propres aux programmes informatiques de la plateforme. Nous omettons ici les messages effacés ou les profils supprimés. Nous interrogeons moins la question de l'exhaustivité de ces résultats que celle de l'agencement et de la performativité qu'engendrent la mise en interface des résultats du moteur. De plus, d'un point de vue méthodologique, nous appuyer sur ces résultats suppose, pour un corpus de plusieurs milliers de tweets, un travail ethnographique (inscription de chaque élément du message dans une grille, impression d'écran) chronophage, ne permettant pas un croisement avec d'autres données en provenance d'autres plateformes. Comment pouvons-nous « brutifier » les données fortement mises en forme par Twitter, afin de réaliser des analyses reproductibles d'une plateforme à une autre ?

L'art de bricoler pour extraire des traces

Ni la documentation de l'API, ni la littérature scientifique ou professionnelle existante sur la question, ne procuraient une méthodologie adaptée à nos objectifs. Nous avons donc opéré un bricolage tant méthodologique que technique en agençant des bribes de méthodes, d'usages, de pratiques, d'outils et de code informatique afin d'obtenir des données exploitables.

Sans prôner une coupure épistémologique entre « science » et « bricolage » (Latour, 1993), nous souhaitons souligner ici un bricolage méthodologique mis en œuvre. Il consiste à *« utiliser de façon créative les matériaux laissés par d'autres projets pour construire de nouveaux artefacts* » sans se reporter à des directives déjà existantes et reproductibles (Rogers, 2012). Cette pratique n'est pas nouvelle pour les recherches qualitatives. Adapter ces méthodes aux contextes et aux acteurs permet de mieux saisir le sens de leurs actions (Denzin et Lincoln, 1999). Ce bricolage peut aussi être

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

conceptuel lors de travaux interdisciplinaires (Kincheloe, 2001). Souvent, les travaux sur le bricolage méthodologique s'intéressent au « mix » de méthodes qualitatives (Waechter-Larrondo, 2005) là où notre bricolage s'est intéressé à des approches quali-quantitatives. En l'occurrence nous avons « aspiré » de manière semi-automatisé dans les résultats du moteur des éléments des tweets (contenu, date, volume de retweets, etc.). Dans le navigateur Google Chrome, nous avons eu recours à un « plug-in » (*Scrapper*). Cet outil permet de définir les « chemins » (//div/a/span/b) dans le code HTML/XML nécessaire pour récupérer, par exemple les URL présentes dans les messages ou le nom d'un compte.

Le recours au langage XPath pour obtenir les éléments voulus a ensuite été automatisé via l'outil Google Spreadsheet. Ce travail d'aspiration a nécessité de redéfinir notre méthodologie en braconnant face à la stratégie de Twitter interdisant ces pratiques de *scrapping* et en agençant des savoirs issus de forums, de lectures de tutoriels ou de visionnage de vidéos.

L'ensemble de la méthode d'extraction des données de ce projet s'est ainsi adapté aux contraintes techniques, commerciales, tout en se devant d'être reproductible sur plusieurs corpus liés à notre enquête. Pour résumer, notre méthode a assemblé des approches techniques (pour un traitement quantitatif) et méthodologiques (pour un traitement qualitatif) qui ont autant participées à créer notre corpus qu'à le révéler, à extraire des données qu'à les « brutifier ». La performativité de notre bricolage méthodologique entre ainsi en résonance avec les contraintes techniques, commerciales et d'usages imposées par les plateformes.

« DIY is law » ou l'obligation de transposer nos méthodes

Notre volonté a été de développer une méthode d'analyse longitudinale d'un ensemble de données liées au financement participatif selon leur répartition spatiale mais également de mettre en récit des données par des entretiens et une observation des

5ème colloque international « Réseaux Sociaux, Traces Numériques et Communication Electronique », 6-8 juin 2018, Le Havre.

espaces de promotion et d'interaction sur le web. Nous avons pu constater que nos méthodes d'analyse des « traces-artefacts » sont loin d'être reproductibles entre les plateformes et dans le temps. Chacune des plateformes pouvant modifier ces conditions d'utilisation des API ou ces traitements de résultats à tout moment. Les dispositifs d'enquête récurrents comme sur les pratiques culturelles des français ne sont donc plus reproductibles dans le cadre d'analyse des traces numériques.

Toute une ingénierie de projet doit alors se mettre en place, pour laquelle les chercheurs en Sciences de l'information et de la Communication sont peu (in)formés: obtention d'une infrastructure matérielle et logicielle; déclaration CNIL; contrat de confidentialité et de propriété intellectuelle. L'analyse multiplateforme des traces nous oblige à recombiner les traces collectées et qui sont produites par des acteurs hétéroclites : industriels, plateformes, autorité publique et site web participatif[6]. Ces traces ne produisent pas des « mobiles immuables » (Latour, 1993) même si nous pouvons recombiner les traces extraites de plusieurs plateformes. Elles nous obligent à transposer nos méthodes d'analyse qualitative afin de comprendre comment les informations circulent entre les plateformes.

Pour Lessig « *code is law* » (2000). Aussi nous nous sommes adaptés aux restrictions des plateformes et aux objectifs de leurs concepteurs et de ceux qui codent. Nous avons dû faire preuve de créativité, risquant de perdre en fiabilité, afin de braconner leurs règles. Ainsi, nous pourrions dire que « *Do It Yourself is law* ». L'extraction de données hors des scénarios d'usages édictés par les plateformes et la transposition d'une analyse endogène, *i.e.* propre à un dispositif et située sur un corpus déterminé, à d'autres plateformes supposent un travail à la fois créatif et invisible pour produire des données manipulables et *in fine* des informations reflétant autre chose que le regard

[6] Pour plus de détails, consulter https://ternumeric.hypotheses.org/197.

imposé par les dispositifs. Ce travail invisible a été souligné par Jeanneret (2011, p. 69) au sujet des traces. « *À travers la masse des traitements qui ont peu à peu contraint les documents de toutes natures à alimenter un observatoire quantitatif, l'hypothèse que tel ou tel trait dans l'écriture (la récurrence d'un terme, la présence d'un lien, la marque d'un geste de consultation) donne accès à la présence d'une réalité sociale est devenue aussi impérieuse qu'invisible.* »

Comme les infrastructures techniques cachent certains pans de leur fonctionnement, les infrastructures de recherche (que l'on nomme trop souvent « projets ») nécessaires aux recherches sur le numérique (car elles associent des moyens techniques, organisationnels et humains) mettent en invisibilité une partie du travail des chercheurs pour produire de l'information scientifique. Ainsi que le mettent en relief Denis et Pontille dans un autre contexte (2012) *« Plutôt que de prétendre que la « véritable » nature de l'information réside dans des formes incertaines en coulisses, l'enjeu est de comprendre comment et au nom de quoi celles-ci sont effacées dans la version stabilisée et solidifiée. C'est en documentant dans le même mouvement le travail invisible de l'information, la multiplicité des matériaux qu'il implique et les formes de son invisibilisation que nous pourrons alimenter une discussion constructive du modèle de l'information désincarnée ».* Il nous semble alors impératif de documenter ces bricolages méthodologiques, ce que nous proposons de faire ici pour notre projet.

Références bibliographiques

Alloing, Camille (2014). Vers une approche instrumentale de l'identité numérique : les attributs identitaires comme structuration de l'environnement informationnel ? In Pinte, J-P. *Enseignement, préservation et diffusion des identités numériques*, Paris : Hermes Lavoisier, pp.39-68.

boyd, Danah, Crawford, Kate (2012). 6 provocations for Big Data, trad. Allard Laurence, Grosdemouge Pierre, Pailler Fred, In Mounier, Pierre, *Read/Write Book 2*, Marseille: Open édition Press, pp. 197-219

Dagiral, Éric, Peerbaye, Ashveen (2013). Voir pour savoir. Concevoir et partager des « vues » à travers une base de données médicales. Réseaux, 2 n°178-179, pp. 163–196.

De Certeau, Michel (1990). *L'invention du quotidien*. t.1. Paris : Gallimard, 1990, 352 p.

Denis, Jérôme, Goëta, Samuel (2013). La fabrique des données brutes. Le travail en coulisses de l'open data. *Journée d'études SACRED* - Paris.

Denis, Jérôme, Pontille, David (2012). Signalétique du métro et politique de l'attention. *Sciences de la Société*, Presses universitaires du Midi, pp.21-39.

Denzin, Norman K., & Lincoln, Yvonna. S. (Eds.). (1999).The SAGE handbook of qualitative research (3rd ed.). Thousand Oaks, CA: Sage Publications.

Helmond, Anne (2015). The Platformization of the Web: Making Web Data Platform Ready. *Social Media + Society*.

Jeanneret, Yves (2011). Complexité de la notion de trace. De la traque au tracé. In Galinon-Mélénec Béatrice, *L'Homme trace. Perspectives anthropologiques des traces contemporaines*. Paris: CNRS Éditions, pp. 59-86.

Kincheloe, Joe L. (2001). Describing the bricolage: Conceptualizing a new rigor in qualitative research. *Qualitative inquiry*, *7*, n°6, pp. 679-692.

Latour, Bruno (1993). Petite leçon de sociologie des sciences. Paris : La Découverte, 252 p.

Le Béchec, Mariannig, Alloing, Camille (2016), Les humanités numériques pour repenser les catégories d'analyse. Le cas du « territoire numérique de marques », in *RFSIC*, 8.

Le Béchec, Mariannig, Dejean, Sylvain, Alloing, Camille, Meric, Jérôme (2017). Le financement participatif des projets culturels et ses petits mondes. 2017. 〈halshs-01508423〉

Lessig, Laurence (2000). Code: and other laws of cyberspace. Version 2.0. New-York: Basic Books, 432 p.

Marres, Noortje, Gerlitz, Carolin (2016). Interface methods: renegotiating relations between digital social research, STS and sociology. *The Sociological Review*, n° 64, pp. 21-46.

Rogers, Matt (2012). Contextualizing theories and practices of bricolage research. *The qualitative report*, *17*, n°48, pp. 1-17.

Waechter-Larrondo, Virginie (2005). Plaidoyer pour le bricolage et l'enracinement des méthodes d'enquête dans le terrain : l'exemple d'une recherche sur le changement dans les services publics locaux. *Bulletin de méthodologie sociologique*, 88, pp. 31-60.