L'éditorialisation des données aux bornes des API : Enjeux et perspectives pour une analyse empirique

The Editorialization of the data at APIs' bounds: Issues and perspectives for an empirical analytic

Editorialización de datos en los terminales de la API: Temas y perspectivas para el análisis empírico

Article inédit, mis en ligne le 15 novembre 2018.

Cette réflexion est associée à la conduite de projets menés dans le cadre du WP4 du Grenoble Alpes Data Institute, supporté par l'Agence Nationale pour la Recherche - Investissements d'avenir (ANR-15-IDEX-02).

Jean-Marc Francony

Jean-Marc Francony est Maître de conférences en sciences de l'information-communication à l'Université Grenoble Alpes. Il est membre permanent du laboratoire PACTE en sciences sociales. Ses recherches portent sur les dispositifs info-communicationnels numériques, l'analyse des pratiques et des usages sociaux de l'Internet, ainsi que sur le développement de méthodes digitales dans un contexte de données massives.

Plan de l'article

Résumé

Les interfaces publiques d'accès aux données contribuent à la définition d'un écosystème de services pour les plateformes du Web 2.0. La publicisation de ces données impose pour ces acteurs de l'économie numérique de concevoir une ouverture graduelle monétisable et d'opérer une réduction informationnelle qui s'apparente à un processus d'éditorialisation. La compréhension de ces mécanismes est essentielle dans la conduite de projets d'analyse de données et de recherches empiriques ou pour nourrir les questionnements critiques, méthodologiques et épistémologiques en SIC.

Mots clés

API, Twitter, *Internet Studies*.

Abstract

Applicative public interface to data are used by Web 2.0's platforms to create an ecosystem of services. For these actors of the digital economy, Data publication involve to conceive a monetisation and a gradual opening and to carry out informational reduction which is similar to an editorial process. The understanding of these mechanisms is essential in the conduct of data analysis and empirical research or to feed critical, methodological and epistemologicial questions in ICS.

Keywords

API, Twitter, Internet Studies.

Resumen

Las interfaces públicas de acceso a datos contribuyen a definir un ecosistema de servicios para las plataformas Web 2.0. La publicación de estos datos requiere que estos actores de la economía digital diseñen una apertura gradual y monetizable y realicen una reducción informativa que se asemeje a un proceso de editorialización. La comprensión de estos mecanismos es esencial para llevar a cabo proyectos analíticos e investigaciones empíricas o para alimentar cuestiones críticas, metodológicas y epistemológicas en el SIC.

Palabras clave

API, Twitter, Estudios en Internet

Introduction

Comme d'autres disciplines des sciences humaines et sociales, les sciences de l'information et de la communication (SIC) sont confrontées aux enjeux des flux de données massives du Web, et tout particulièrement ceux associés aux réseaux sociaux. Si la prise en compte de ces flux s'impose en tant qu'objet et ne fait pas débat, il en va tout autrement de leur exploitation et de leur interprétation (Bigot et al., 2016), (Paquienséguy et al., 2017). Les questions soulevées renvoient à des considérations méthodologiques mais aussi épistémologiques.

La genèse de l'expression Web 2.0 et les difficultés à la justifier rétrospectivement fonde une critique légitime de son usage scientifique (Bouquillion et Matthews, 2010). De plus, le terme Web 2.0 a étendu la définition du Web à l'ensemble des plateformes dites réseaux sociaux numériques qui ont émergé depuis 2000. S'il ne peut être question d'une notion caractérisant une pratique sociale ou un

modèle économique, ce jalon historique reste, à nos yeux, pertinent dans l'évolution des techniques de l'Internet. En effet, du point de vue de l'architecture informatique, les années suivantes voient un modèle de plateforme de services s'imposer comme standard du Web[1].

Dans le même temps, les flux de données rendus disponibles par les opérateurs du Web acquièrent de la valeur. Comprendre les modalités de leur fabrication et de leur publicisation est indispensable en amont de leur exploitation. La rétro-ingénierie informationnelle apporte un éclairage sur ces modalités. En suivant cette approche, nous nous intéressons aux flux de données publicisés dans les API de Twitter. L'objectif est à la fois de comprendre le processus de fabrication et ainsi la nature des données disponibles et leur portée, mais également de formuler des hypothèses sur les logiques de leur publication.

Les enjeux des API

Tim O'Reilly (O'Reilly, 2005) a eu le mérite d'énoncer les caractéristiques fonctionnelles et informationnelles désormais dominantes de ces plateformes de services. Les préconisations associées à cette conception encouragent les formes ouvertes et collaboratives de développements informatiques. La mise en œuvre d'interfaces de programmation applicatives ou API (*Applications Programming Interface*) est une méthode d'ouverture par les données visant le développement de services annexes et ainsi la croissance d'un écosystème périphérique à la plateforme[2]. L'avènement d'un Internet des Objets (*IoT,* pour *Internet of Things*, aussi identifié comme Web 4.0) conforte ces choix structurels et impose le recours aux API dans le développement de systèmes numériques complexes pour en garantir l'interopérabilité (Institut Montaigne, 2015).

Bien que l'information mise à disposition aux bornes des API se doive d'être le reflet de leur activité, il ne s'agit cependant pas, pour ces entreprises que sont les plateformes de services, de rendre accessibles les éléments d'un avantage concurrentiel sans contrepartie. Ainsi voit-on de plus en plus le développement des API publiques se décliner suivant différents modèles d'accessibilité plus ou moins étendue en fonction d'une échelle de rémunération associée. L'accès gratuit devient un mode standard dont les restrictions qualitatives et quantitatives sur le flux sont justifiées par l'existence de modes premiums payants offrant graduellement un accès plus performant.

Annoncées pour le printemps 2018, les nouvelles perspectives réglementaires européennes sur la protection des données personnelles (GDPR) incitent les acteurs du Web 2.0 à réduire toujours plus l'ouverture sur les données qu'ils élaborent, tout au moins à en maîtriser davantage leur diffusion. En effet, pour ces plateformes, il est indispensable d'intégrer les interactions possibles entre les données publicisées via les API avec celles issues des nouvelles possibilités de valorisation découlant du principe de portabilité des données personnelles (article 20-2).

Dans ce contexte actualisé, nous formulons l'hypothèse que le processus de réduction informationnelle opérée aux bornes des API publiques s'apparente à un processus d'éditorialisation qui en fixe la valeur d'usage et oriente le développement de l'écosystème de services.

L'accès aux données d'Internet relatives aux activités humaines constitue un enjeu de premier plan dans la compréhension de l'évolution sociétale ou des pratiques sociales des dispositifs numériques. L'usage intensif d'objets connectés au Web, la complexité croissante des interactions médiatisées

. . . . . .

1 Les travaux sur l'architecture orientée service (SOA) datent du début des années 2000.

[2] On évoque pour Twitter un trafic 10 fois supérieur via ses API que via sa plateforme Web. (http://avc.com/2007/09/biz-stone-on-re/)

rendent nécessaire l'accès aux données publicisées par les plateformes de services. De nombreux acteurs publics et privés s'emploient ainsi à utiliser les API publiques pour collecter et analyser les données. Cependant, la généralisation de la monétisation des flux crée des paliers entre l'accès gratuit et les modalités payantes dont les conséquences qualitatives sur les analyses méritent d'être soulignées.

L'exemple de Twitter est de ce point de vue emblématique. Cette plateforme du Web social est probablement l'une des plus étudiées. L'instantanéité de ses publications fait de Twitter un outil privilégié d'alerte lors de la survenue d'événements, ou de signalement lorsque de nouvelles publications sont en ligne. Cette caractéristique se double de l'avantage qu'offrent les API de Twitter publiques d'accéder en masse aux représentations numériques très structurées de tweets qui concentrent pour chacune d'elles une information qu'il serait fastidieux de collecter depuis la page Web de comptes Twitter. L'accessibilité et la diversité des données ainsi structurées éveillent l'intérêt d'analystes souhaitant produire des services ou de la connaissance à partir des flux de Tweets ou de leurs réseaux de diffusion.

En partant de l'exemple de Twitter, dont les quelque 500 millions de messages instantanés journaliers suscitent une convoitise, nous proposons de formaliser les restrictions informationnelles issues de l'éditorialisation des données, et nous envisageons les conséquences de ce processus du point de vue général de l'économie des données mais également du point de vue méthodologique et épistémologique des sciences de l'information et de la communication.

L'analyse des flux de données : l'exemple de Twitter

Au travers des conditions générales d'usages (CGU), de la définition fonctionnelle des API ou enfin de la construction du flux informationnel, le processus éditorial définit un cadre normatif pour l'analyse. Pour la recherche, la connaissance de ce cadre est une nécessité méthodologique. Elle permet d'évaluer l'étendue, la pertinence et la portée heuristique des collections envisagées. Dans le cas de Twitter, c'est la restriction en volume du flux délivré qui est généralement évoquée comme limitation de l'API gratuite. Mais la conséquence de cette limitation, établie à 1% du flux total, n'en reste pas moins floue pour la majorité des travaux de recherche (Morstatter et al., 2014). En novembre 2017, Twitter a récemment réaménagé son modèle économique et a profondément remanié l'accès à ses interfaces de données. L'offre se trouve enrichie d'accès historiques, géolocalisés ou encore d'informations volumétriques, selon des opérateurs premiums dont les performances, calibrées selon une grille tarifaire mensuelle[3], sont susceptibles de répondre aux besoins qualitatifs et quantitatifs d'entreprises émergentes.

Les modalités d'accès

Collecter des informations aux bornes des API implique l'utilisation de logiciels lui permettant de communiquer avec les serveurs supportant les API et de leur faire exécuter les fonctionnalités requises. Il n'existe pas véritablement de solution clef en main permettant de réaliser cet ensemble d'opérations simplement. On trouve majoritairement des modules logiciels jouant le rôle de connecteur fonctionnel avec un langage de programmation de plus ou moins haut niveau.

[3] Le plancher premium est fixé à environ 150$/mois et peut être multiplié par dix pour l'offre entreprise. Cette offre complète celle antérieure, beaucoup plus élevée, accessible auprès d'opérateurs chargés de la valorisation des flux de données.

L'investissement technique nécessaire pour les utiliser est faible lorsqu'il s'inscrit dans l'effort plus soutenu qu'implique la maîtrise des outils analytiques contemporains comme R par exemple.

Dans ce contexte, l'accès aux flux de données implique une démarche équivalente à celle d'un développeur logiciel qui doit d'abord déclarer son projet de développement auprès de la plateforme (Twitter) afin d'obtenir une clef d'authentification applicative (*consumer key*). Cette déclaration préalable est de nature contractuelle et engage le développeur à suivre les conditions générales de la politique de développement de Twitter[4]. Cette contractualisation vient en complément d'un autre engagement contractuel associé aux conditions générales d'usages (CGU) d'un compte Twitter. Il est fait référence à ce compte dans l'identification nécessaire via des clefs d'utilisation (*access token*) pour les services de Twitter. Ces différents contrats établissent la responsabilité personnelle (*accountability*) au regard de la balance entre les droits attribués et les devoirs contractés.

Parmi les devoirs du développeur figure le respect des modalités de sollicitation de l'API. Celles-ci font état des contrôles ascendants (client>serveur) ainsi que des contrôles descendants (serveur>client) suivant la nature des requêtes ou de l'état des services sollicités. Ces restrictions impactent notamment l'API SEARCH dont le fonctionnement est réglé en fréquence de requêtes et en volume de réponses. Pour l'API STREAM, les restrictions portent plutôt sur la complexité du filtre de flux.

Les différentes CGU définissent en partie les restrictions informationnelles opérées par Twitter. Au-delà de ces restrictions qui sont clairement énoncées et monitorées par la plateforme, d'autres contraintes encadrent les productions informationnelles disponibles aux bornes des API.

Les API RESTful

En informatique, le concept d'interface est associé à l'organisation logique des traitements. Les interfaces articulent des fonctionnalités qui se distinguent du fait de leur nature ou des modalités de leur mise en œuvre. Le principe des interfaces s'est imposé dans le génie logiciel comme le moyen de maîtriser la complexité et d'assurer des développements informatiques modulaires et indépendants.

L'acronyme REST (*REpresentational State Transfer*) qualifie des interfaces donnant accès à des représentations associées aux entités manipulées (ou objets) dans l'application, selon un cahier des charges spécifique tendant à standardiser la production. Les directives REST reprennent les principales caractéristiques de la définition du Web dont en particulier le protocole de communication HTTP pour le transport des données, l'architecture client-serveur et l'absence de mémorisation d'états de sessions. Outre l'uniformisation du Web, les objectifs associés à la norme REST portent sur la qualité des performances (côté serveur) quel que soit le dimensionnement (*scalability*) de la demande.

On associe aussi le terme REST à l'adjectif RESTful (paisible) pour caractériser la conformité des modalités d'accès aux données de l'application suivant ce standard. En effet, sa mise en œuvre favorise l'interopérabilité applicative et uniformise le travail de développement des programmeurs, l'allégeant du même coup. L'appellation RESTful est de ce fait revendiquée par les plateformes comme un gage de qualité sans pour autant se conformer entièrement au standard.

Dans le contexte du développement des plateformes de services, la caractérisation REST s'applique aux API développées, contribuant ainsi à l'émergence d'une norme d'usage. Le choix d'un développement de type REST permet de séparer clairement les rôles dans l'accomplissement du

. . . . . .

[4] https://developer.twitter.com/en/developer-terms/agreement-and-policy, consulté le 21/05/2018.

service en se rapportant au modèle client/serveur. En particulier, cela permet de déléguer au client le soin de contextualiser ses requêtes et de mémoriser les éléments de session pertinents, ce qui allège d'autant la charge du serveur.

Twitter comporte un ensemble d'API (plus d'une dizaine) dont l'étendue et les spécificités évoluent en fonction de l'offre de services et de la politique d'ouverture des données. L'abandon récent de la restriction des 140 caractères du corps de message pour aller vers un format documentaire multimédia plus étoffé n'est pas la moindre de ces évolutions. Dans les faits, la limitation est maintenue pour l'affichage du texte mais ne l'est plus pour l'encodage textuel du message. Les textes peuvent désormais être plus longs mais masqués. Les informations structurelles du message, comme par exemple les comptes mentionnés, ou les URLs tweetées ne sont plus comptabilisées. Cette définition étendue s'aligne avec celle du billet d'information enrichie imposée par la concurrence, notamment de Facebook. Elle vise également la facilité de réponse pour favoriser le régime conversationnel des échanges et en conséquence la production de contenus.

Cette modification affecte le nombre et la nature des objets de Twitter. La structure informationnelle correspondante s'enrichit notamment de relations structurelles inédites. Il en découle de nouvelles possibilités de représenter et d'accéder aux informations associées à ces entités. Ces possibilités se traduisent par des représentations structurées qui se redéploient suivant une organisation d'interfaces elle-même renouvelée. Toutefois, la rétrocompatibilité qui est nécessaire avec les développements antérieurs de la plateforme Twitter et de son écosystème atténue la portée de ces évolutions. Toutes ces API ne respectent pas strictement la définition REST mais la plupart d'entre elles en sont cependant dérivées.

Les représentations publiques gratuites

Compte tenu des enjeux portés par les données, nous distinguons formellement deux systèmes de représentations. Le premier répond aux logiques du système d'information qui supporte l'activité de la plateforme et ses développements stratégiques. Ce système de représentation (interne) est confidentiel. Même si une partie des données qu'il organise fait l'objet de publicisation, celle-ci mobilise un système de représentation (externe) distinct. Ce deuxième système répond aux nécessités d'une communication maîtrisée respectant les contraintes réglementaires, éthiques et stratégiques. Dans le cas de Twitter, la définition et l'organisation des API sont étroitement liées à ce système. En effet, la décomposition et les fonctionnalités des API reprennent les principes d'une conception centrée objet. Dans le cas présent, les représentations numériques mises en œuvre dans ces interfaces formalisent les entités conceptuelles et les actions associées aux modèles d'information et d'interaction adoptés par cette plateforme.

C'est ainsi que la représentation d'un tweet (STATUS) contient les représentations d'entités associées à l'auteur (USER), à des listes de composants (HASHTAG, MEDIA, URL, SYMBOL) et enfin, à l'extension du message qui correspond soit au tweet originel (STATUS) dans le cas d'un retweet soit à une liste de composants supplémentaires dans le cas d'un tweet long. L'attribut QUOTED_STATUS permet de repérer les cas de citations commentées, c'est-à-dire ajoutant une contribution à ce qui sinon n'aurait été qu'une rediffusion (RETWEETED_STATUS).

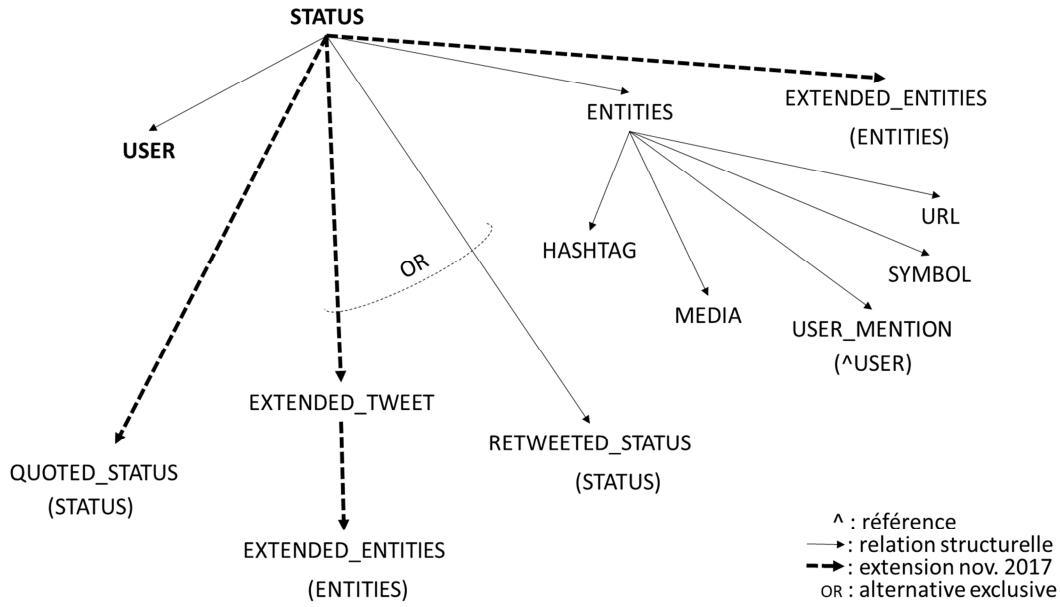

Figure 1. Schéma structurel externe d'un Tweet[5].

Seuls les deux représentations des STATUS et des USER comportent une clef d'identification interne qui en assure l'unicité et constitue le moyen de formuler des requêtes pour d'autres API connexes.

Les clefs d'identification fournies rendent compte de l'existence des deux entités STATUS et USER dans le système représentationnel. Leur présence conjointe dans la représentation du message rend probable l'existence formelle de la relation éditoriale d'attribution (*author*) dans le système de représentation interne. L'inscription du message originel dans le cas d'une rediffusion permet en outre de faire exister formellement deux relations orientées entre entités USER : une relation de rediffusion (retweet) entre auteurs (rediffusé et rediffusant) ainsi qu'une relation de mentionnement (*mention*) entre l'auteur et chacun des comptes identifiés dans le message.

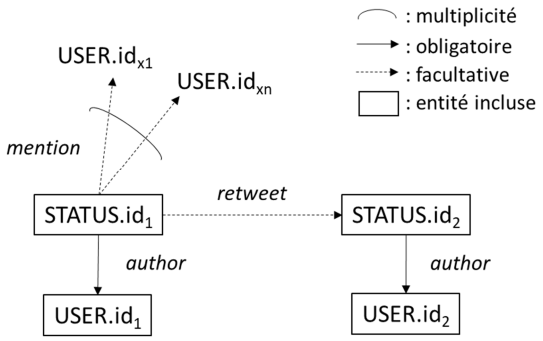

Figure 2. Schéma relationnel externe d'un Tweet

[5] La branche d'extensions (EXTENDED_TWEET) correspond à la sortie des restrictions de publications antérieures, en maintenant la cohérence avec le modèle historique.

Ces différents éléments déterminent l'ensemble des calculs que l'on peut opérer à partir de la capture d'un flux de Twitter. Il est ainsi formellement possible de rechercher dans ce flux et sans ambiguïté, l'ensemble des publications d'un abonné, d'établir l'existence de relations éditoriales entre auteurs et ainsi d'extrapoler des mesures d'intensité de flux ou de liaison entre ces acteurs, sous couvert d'hypothèses portant sur l'échantillonnage aux bornes des API. Cette unicité d'accès garantit en outre l'interopérabilité des services et des développements dans l'écosystème applicatif. Twitter privilégie ainsi l'exploitation des relations éditoriales mettant l'accent de cette manière sur une forme de réseautage que l'on peut qualifier de circonstanciel. Il s'agit en effet, de réseaux tributaires de la capture d'un flux qui ne vaut que dans le contexte spécifique de sa production et dans la durée des faits qu'il relate. L'analyse des graphes de rediffusion ne doit pas oublier qu'il s'agit de configurations dynamiques d'acteurs éventuellement éphémères. De ce point de vue, les représentations résultantes se distinguent de celles des réseaux déclaratifs de suivi (*follow*) dont la stabilité est plus grande mais dont le besoin de connaissance nécessite d'interroger une API distincte. Compte tenu des règles de cumul appliquées sur les sollicitations authentifiées des API, cette double interrogation ne peut pas être conduite simultanément avec une seule machine sans risque de discontinuité de flux. Ce jeu de contraintes techniques témoigne de la cohérence des limitations imposées et de l'éclatement fonctionnel entre les API de Twitter. Cette distribution informationnelle dans le système d'information publique a comme conséquence de réduire la possibilité d'analyse au fil de l'eau des réseaux d'acteurs et des flux informationnels. Twitter préserve ainsi, et pour son compte, la valeur événementielle du flux et l'offre de services qui en découle.

L'éditorialisation de flux

L'adaptation informationnelle réalisée par les opérateurs de plateformes du Web 2.0 s'inscrit dans une stratégie globale de mise en visibilité de leurs services et de contrôle des ressources disponibles. Elle constitue désormais ce que l'on peut considérer comme un processus d'éditorialisation (Lipsyc et Ihadjadene, 2013) généralisé à l'ensemble de ces ressources numériques. Ce processus s'applique en effet, non seulement à l'ensemble des publications sur le Web mais aussi dans les flux de données publicisés depuis les API. Le rôle de ces interfaces spécialisées n'est donc plus strictement de faciliter l'accès aux représentations numériques internes mais d'en contrôler l'accès, la diffusion et l'usage possible. La sélection des informations publicisées constitue la première étape de ce contrôle. La deuxième étape consiste à structurer le système représentationnel externe en fonction des objectifs de publicisation et des exigences de contrôle sur le potentiel heuristique des représentations produites.

L'alignement métaphorique

Pour Twitter, la logique de contrôle sur les contenus est cohérente avec un système métaphorique fondé sur les notions d'événement et de flux qui sous-tendent son fonctionnement. Dans ce contexte, la publication d'un message est assimilée à un événement éditorial. Ce dernier correspond à un événement qui survient puis s'estompe au regard de nouveaux événements plus « prégnants ». La prégnance d'un événement est définie par l'intensité relative de sa rediffusion (retweet) au sein de l'espace de publication courant. L'intérêt de ces métaphores est d'assurer la cohérence avec le modèle des médias de flux. Elles donnent un sens aux notions d'actualité (*news*) et de tendance (*trend*) qui concourent au succès de Twitter dans l'espace médiatique. Elles justifient également la fugacité des événements de publication et donc instaure la durée de vie des messages.

L'analogie avec les médias mise en œuvre par Twitter ne se limite pas aux conditions d'usages des services de publication proposés. Elle se prolonge dans la définition de ses API et des logiques de publicisation des données. À ce niveau, l'effacement progressif des données est réalisé dans le

système représentationnel au moyen d'un historique de publication auquel on ne peut accéder que de manière antéchronologique et pour un volume limité de messages. L'accumulation progressive de messages publiés dans l'historique conduit à un effacement d'autant plus rapide que le flux de publications est important2. Selon les modalités d'accès choisies, la portée historique d'une semaine (gratuit) est repoussée aux 30 derniers jours (premium) voire sans limite pour le mode « entreprise ».

Le mécanisme de rediffusion (retweet) est cohérent avec une communication de proche en proche portée par des réseaux de socialité. L'objectif associé est à la fois d'organiser la rareté de l'information et d'assurer la promotion d'un message en maintenant son actualité. Sa traduction fonctionnelle est d'assurer la rémanence des contenus informationnels du message originel dans l'historique de publications mais pour une durée qui ne dépasse que très rarement une semaine.

La mise en cohérence métaphorique naturalise en retour les restrictions d'accès aux données de publication. Le modèle éditorial des médias justifie, par exemple, l'identification de l'auteur originel d'un message rediffusé et non des intermédiaires qui ont participé à sa diffusion. Cette référence unique et légitime à la source est inspirée des bonnes pratiques de l'édition. Elle interdit en conséquence l'accès aux chaînes de rediffusion et de ce fait, à la connaissance des réseaux de diffusion effectifs (*followers*). De cette manière, l'identification des acteurs clefs de l'information est beaucoup plus difficile à appréhender, ce qui a des répercussions dans la mise en œuvre de modèles de communication.

Les réductions représentationnelles

D'autres formes de restrictions interviennent en complément des limitations de flux ; de nature représentationnelle, elles consistent à traduire sous une forme moins riche les représentations internes disponibles. Cet effacement n'est effectif qu'au niveau représentationnel externe. L'ensemble des messages (hors embargo) peuvent être retrouvés à partir de services payants assurés par des prestataires chargés de la monétisation des données de Twitter (Gnip, etc.). Nous distinguons principalement trois méthodes :

- le référencement revient à identifier les entités par un code interne plutôt que de fournir leurs représentations complètes. Il s'agit d'un moyen terme qui fait peser sur le client un coût supplémentaire (requête) s'il souhaite accéder à cette représentation sachant que l'interaction avec une API est aussi contrôlée. Dans tous les cas, la résolution référentielle désynchronise les représentations qui peuvent désigner deux états distincts d'une même entité. C'est le cas avec les comptes mentionnés mais aussi avec les URL qui pointent des éléments éventuellement externes au système de représentation interne de Twitter. Dès lors, les corrélations ne sont plus fondées temporellement ;
- l'écrasement consiste à restituer une forme de surface non structurée qui ne permet pas non plus d'établir une référence à une entité représentationnelle. Les composants d'un message (HASHTAG, etc.) rendent compte de limites qui tiennent autant au développement de la plateforme et de ses services qu'à la volonté de ne pas fournir les moyens d'une captation de valeur trop évidente. Si ces entités peuvent éventuellement avoir une existence formelle dans le système de représentation interne, elles n'en ont probablement pas dans le système externe. L'effort de représentation est ainsi reporté côté client et réduit la portée des services proposés.
- la synthèse est réalisée au moyen de catégories ou de quantifications produites à la place de représentations dénombrables d'entités. La quantification peut également s'accompagner d'un échelonnement des valeurs qui fixe une granularité de l'information fournie. Il est à noter ici que Twitter donne la valeur exacte des compteurs (*followers*, etc.) à la différence de

LinkedIn qui ne fournit pas de manière précise le nombre de relations personnelles au-delà de 500 contacts.

La distribution fonctionnelle

Comme nous avons pu le voir avec Twitter, les opérations précédentes s'appliquent sur une sélection d'attributs déterminée par le choix d'une distribution fonctionnelle. Il s'agit de faire en sorte qu'aucun appariement de collections de données ne permette, en quantité et en qualité, de produire une information estimée stratégique pour l'opérateur de la plateforme. Pour empêcher les appariements (ou croisements de données), il suffit que les représentations soient disjointes, c'est-à-dire qu'il n'y ait pas d'identifiant unique permettant une mise en correspondance immédiate. Si toutefois la jointure est inévitable, celle-ci doit alors avoir une portée limitée ou induire un coût de réalisation dissuasif ou monétisable.

Dans le cas de Twitter, ce contrôle global est réalisé en monitorant les transactions authentifiées. Le contrôle en fréquence et en volume de transactions allonge les délais de calcul, atténuant du même coup leur portée.

Conclusion

Certes moins privative que d'autres plateformes (Facebook, LinkedIn, etc.), l'évolution de la politique de publicisation des données de Twitter témoigne d'une économie numérique où la valeur s'établit désormais sur l'abondance des données.

Pour les plateformes de service, la publicisation des données a pour objectif de constituer un écosystème et de créer des opportunités de marchés. Cette ouverture des données ne peut s'envisager sans contrôle ni contrepartie de la valeur ainsi cédée. La mise en œuvre de modalités payantes contribue à cette régulation. Le ticket d'entrée élevé qui en découle établit de nouvelles frontières financières entre les tiers d'une économie numérique. On identifie donc, d'un côté les acteurs dotés de financements solides, porteurs de projets et disposant des moyens d'une exploitation massive de flux de données et de l'autre, des acteurs à la recherche d'opportunités, engagés dans une démarche exploratoire. Cette partition vaut également pour le monde académique qui ne fait pas exception, malgré l'évolution des cadres légaux[6]. Dans un contexte règlementaire prioritairement favorable à l'économie numérique, contourner l'écueil des coûts impose des regroupements institutionnels afin d'assumer le cofinancement d'études ou d'atteindre la masse critique nécessaire pour être éligible lors d'appels à projets. Le changement d'échelle imposé dans la conduite de tels projets fait écho aux processus de rationalisation porté par les institutions scientifiques. Toutefois, les efforts de coordination scientifique restent d'un effet limité puisqu'au regard des CGU fixées par Twitter, il ne peut être question d'ouvrir[7] des collections de Tweets sans s'exposer à des poursuites. Relâcher cette contrainte devient une nécessité scientifique et citoyenne pour laquelle nos communautés scientifiques doivent se mobiliser.

En dépit des restrictions qualitatives ou quantitatives opérées sur les données, les API maintiennent les caractéristiques de flux instantanés massifs, témoignant d'activités et d'interactions connectées à

. . . . . .

[6] https://www.economie.gouv.fr/republique-numerique, https://ec.europa.eu/research/openscience/index.cfm?pg=openaccess

[7] Au sens de l'open data que promeut la loi sur la république numérique. La seule solution jugée satisfaisante par Twitter étant de ne partager que les identifiants de Tweets (Status.id) pour une période inférieure à 30 jours. Pour plus de détail voir le post : https://blog.ldodds.com/2017/05/19/can-you-publish-tweets-as-open-data/ (consulté le 01-06-2018)

l'espace social et à la sphère médiatique. Ainsi, l'accès aux flux de Twitter est une source d'intérêt pour l'étude de phénomènes info-communicationnels, dans la limite bien comprise des mécanismes de publicisation et de leurs conséquences sur les données. La compréhension qualitative des données disponibles est une première étape d'un processus analytique mobilisant des techniques et des méthodes mixtes, à la fois qualitatives et quantitatives. Elle offre également la possibilité d'une lecture critique de l'économie de l'information sous-jacente à ce type d'entreprise du Web. Enfin, le changement d'échelle, conduisant aux données massives constitue un objectif distinct que l'on peut néanmoins situer dans la continuité de cette acculturation progressive aux données de l'Internet.

Références bibliographiques

Bigot, Jean-Edouard, Julliard, Virginie, et Mabi, Clément, (2016), « Humanités numériques et analyse des controverses au regard des SIC. Retour sur une expérience pédagogique », *Revue française des sciences de l'information et de la communication*, n°8.

Bouquillion, Philippe, et Matthews, Jacob Thomas, (2010), *Le Web collaboratif. Mutations des industries de la culture et de la communication*, Grenoble : Presses universitaires de Grenoble.

Institut Montaigne (2015), *Big data et objets connectés - faire de la France un champion de la révolution numérique*, Rapport. En ligne : http://www.institutmontaigne.org/res/files/publications/rapport%20objets%20connecte%CC%81s.pdf consulté le 1 juillet 2018.

Lipsyc, Carole, et Ihadjadène, Madjid, (2013), « Architecture de l'information et éditorialisation. In L'architecture de l'information : un concept opératoire », *Études de communication*, (2), n°41, p. 103-118.

Morstatter, Fred, Pfeffer, Jürgen, et Liu Huan, (2014), « When is it Biased? Assessing the Representativeness of Twitter's Streaming API », in *Proceeding WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web*. Seoul, Korea p. 555-556.

O'Reilly, Tim, (2005), « What is Web 2.0? », in *Online communication and Collaboration* - A Reader Ed. Donelan H., Kear K. And Ramage M. The Open University - Routledge p. 225-234.

Paquienséguy, Françoise et al. (2017), « Manifeste pour un positionnement des sciences de l'information communication (SIC) vis-à-vis des Digital Studies (DS) et autres mutations du Numérique », Sous la direction de Françoise Paquienséguy. *Revue française des sciences de l'information et de la communication*, n°10.