Réponse à la consultation publique de l'Arcom sur l’accès aux données des plateformes en ligne pour la recherche

Nicolas HERVÉ, chercheur en informatique
Institut National de l’Audiovisuel (Ina)
n.herve@ina.fr

22 juillet 2022

Une partie de mes travaux consiste à étudier la propagation des informations entre les différentes sphères médiatiques, y compris les réseaux sociaux. L’objectif est de comprendre et modéliser le fonctionnement de cet écosystème ainsi que les interactions entre ses différents acteurs. Pour cela, dans une approche quantitative, il est nécessaire d’avoir accès aux différents contenus afin de les analyser à l’aide d’algorithmes pour en extraire les informations nécessaires permettant de quantifier les phénomènes que l’on cherche à étudier [6]. La propagation des informations rentre à mon avis dans le champs de l’article 31 du DSA. Les réponses aux questions de la consultation se placent donc dans le cadre de mon expérience sur plusieurs années d’accès aux données de Twitter, Facebook et Youtube. Certaines questions pour lesquelles je n’ai pas d’avis ou de compétence particuliers sont laissées de côté. Ces réponses représentent un point de vue personnel.

A.1. La principale plateforme utilisée est Twitter [1, 2, 3, 4, 5, 7, 8, 9]. La collecte se fait alors très majoritairement via l’API fournie. Pour Facebook [2, 5, 7], l’utilisation de l’API, même si elle est très limitée et de Crowdtangle a en partie permis de récupérer quelques informations. Enfin, pour Youtube [3], s’agissant simplement de la récupération des vidéos, un *scraper* open-source est utilisé.

A.2. Les principales difficultés rencontrées sont de quatre ordres :

- données non disponibles via une API. Sachant qu’une donnée existe et est publique, l’obtenir de façon automatisée n’est pas toujours aussi facile qu’un appel à une API. Il faut alors passer par des solutions de contournement, principalement du *web scraping*, qui ont l’énorme inconvénient d’être plus coûteuses en temps (nécessité de faire un travail de rétro ingénierie du site web et de développer une solution *ad hoc* à chaque fois), plus coûteuses en ressources informatiques (une émulation de navigateur web peut être nécessaire, beaucoup plus gourmande en CPU et mémoire qu’un appel à une API) et de ne pas être pérennes (sensible à tout changement dans la page web de la plateforme).
- limitation de la volumétrie des appels à l’API. On distingue deux types de limitation par unité de temps : nombre d’appels à l’API et/ou quantité d’informations qui peut être obtenue par ce biais. Là encore il peut être nécessaire de contourner les limitations pour pouvoir mener certains travaux de recherche. Cela peut passer par l’utilisation de plusieurs clés d’API, voire à la mise en place de stratégies plus complexes [9].
- disponibilité des données de façon pérenne. La reproductibilité des travaux est un point essentiel de la démarche scientifique et suppose donc des jeux de données non volatiles. Ainsi, par exemple, lorsqu’un algorithme obtient un niveau de performance sur un corpus de tweets, il est important que d’autres chercheurs puissent confirmer ce résultat et s’y comparer. Il est alors impératif que ce corpus soit disponible à l’identique. Les CGU de Twitter ne permettent que de partager les identifiants de ces tweets, à charge ensuite pour chaque chercheur de recapter les données, s’exposant ainsi au problème des tweets effacés par leurs auteurs ou des comptes disparus ou suspendus.
- données non publique. Dans ce cas, il faut essayer d’estimer cette information, avec tous les biais que cela peut induire dans les résultats. C’est typiquement le cas des métriques

d’audience, qui sont importantes pour toutes les études qui ne se penchent par uniquement sur la production des informations en ligne, mais également sur leur réception.

A.3. Je n’ai pas *stricto sensu* abandonné de projet de recherche faute d’accès aux données. En revanche, conscient des limitations, j’ai parfois orienté mes travaux pour faire en sorte de ne pas être confronté à ce problème. Cette forme d’autocensure, pernicieuse, est très présente dans la communauté et a naturellement eu tendance à orienter les chercheurs vers Twitter pour l’étude de la propagation d’informations en ligne du fait de sa politique de plus grande ouverture des données.

Lorsque cela s’est avéré nécessaire, des solutions de contournement ont donc été mises en place pour accéder aux données. Cela demande un effort d’ingénierie continu qui a pour principales conséquences :

- de limiter la possibilité d’effectuer certains travaux de recherche aux seules équipes au sein desquelles les compétences informatiques nécessaires sont disponibles
- quand bien même ces compétences sont disponibles, de limiter les travaux de recherche originaux car le temps dévolu à l’obtention des données ne peut être investi sur d’autres sujets

Ainsi, lorsqu’une information est publique, il y a une forme d’hypocrisie à se retrancher derrière des considérations techniques pour en limiter l’accès aux chercheurs. Les restrictions sur les API sont légitimes pour protéger les infrastructures des plateformes ainsi que la dispersion de leurs données. Mais on a vu qu’il est toujours possible pour une personne motivée de contourner en partie ces restrictions. Alors même que la directive DANUM permet aux chercheurs de traiter ces données, la capacité à investir dans un surcoût technique et humain ne devrait pas être le ticket d’entrée assurant la sélection des équipes auxquelles ces données sont accessibles. Ainsi l’accès à la nouvelle API V2 de Twitter s’est accompagné d’un effort d’ouverture pour les chercheurs. Il est ainsi possible de fouiller dans l’intégralité de l’archive, sans restriction temporelle, ce qui est assurément une grande avancée. En revanche, cette ouverture se double d’une restriction sur la quantité de tweets qu’il est désormais possible de récupérer (10 millions par mois, alors que la limite était de 1 % du volume total avec la V1). Ce type de nouvelle contrainte est adapté pour des projets de recherche qui se concentrent sur la façon dont un sujet particulier et précis est abordé sur Twitter. Il suppose que l’on soit capable de définir *a priori* un ensemble de requêtes pour délimiter ce corpus [1, 3, 7] et que le volume de tweets ne soit pas trop important. En revanche, les études plus systémiques, qui visent à observer globalement Twitter [2, 4, 8, 9], nécessitent un volume beaucoup plus important de données qui soient statistiquement représentatives des échanges en ligne.

C’est un point essentiel de la politique d’accès aux données des plateformes : la constitution même des corpus nécessite parfois la mise en place d’algorithmes de fouille de données pour identifier les informations qu’il est pertinent de récupérer en vue d’une étude. Il n’y a dans ce cas que trois solutions possibles :

- soit les calculs sont effectués directement sur la plateforme avec de fortes contraintes (techniques, capacités de calcul, confidentialité des travaux des chercheurs, …)
- soit les données sont téléchargées en masse et fouillées au sein des laboratoires de recherche
- soit ces projets de recherche ne peuvent tout simplement pas être menés

Je tiens à rappeler que si des difficultés rencontrées avec Twitter sont évoquées dans ce document, les travaux académiques seraient autrement plus utiles et pertinents si l’ensemble des plateformes appliquaient déjà *a minima* une politique équivalente.

Enfin, il faudra également veiller à ce que la mise en place du DSA et du tiers de confiance ne se traduise pas par des contraintes administratives ayant les mêmes effets délétères si elles

provoquaient une surcharge trop importante de travail pour accéder aux données. Remplacer un frein technologique par un frein administratif ne ferait que déplacer le problème.

B.1. Il semble que les types d'acteurs soient déjà définis dans le DSA. Si une marge de manœuvre existe pour élargir cette typologie, alors oui, il est souhaitable de le faire. Ainsi, typiquement, je peux au titre de la directive DANUM fouiller les données des plateformes, mais au titre du DSA je ne suis pas reconnu comme tel (car non rattaché à une université).

B.3. Je pense qu'effectuer les demandes d'accès en passant par un tiers de confiance ne devrait pas être systématique, mais limité aux cas hors norme ou en cas de recours si un refus de la plateforme ne semble pas justifié. Globalement, tout ce qui peut réduire les lourdeurs administratives va dans le bon sens.

C.3. Il est important de laisser la liberté aux chercheurs dans la définition de leurs objets de recherche. Des appels à projets sur certaines thématiques peuvent être pertinents si le régulateur estime qu'un sujet mérite d'être approfondi, et il peut même éventuellement être couplé avec les agences de financement nationales ou européenne pour allouer des fonds sur ces travaux. En revanche il ne doivent clairement pas être le seul moyen de faire une demande.

C.6. Comme toute évaluation scientifique, seuls les pairs doivent avoir une voix au chapitre dans l'évaluation des projets si cette évaluation conditionne l'accès aux données.

C.9. Une charte de bonne conduite à destination des chercheurs paraît un outil intéressant pour se prémunir des risques évoqués.

D.4. Les API, éventuellement dédiées, sans limitation de volume, sont une approche à privilégier.

E.2. Voir réponses A.2. et A.3.

Références

[1] *Étude quantitative de l'intensité médiatique des 6 premiers mois de la pandémie du Covid-19*, N. Hervé, Les Cahiers du journalisme, 2022 (à paraître)

[2] *Social Media and Newsroom Production Decisions*, J. Cagé, N. Hervé, B. Mazoyer, NBER Political Economy, 2021

[3] *Circulation des vidéos de violences policières entre Twitter et la télévision*, N. Hervé, Working paper

[4] *Quelle modélisation de l'espace politique français sur Twitter ?*, N. Hervé, EGC 2021

[5] The Production of Information in an Online World, J. Cagé, N. Hervé, M-L. Viaud, The Review of Economic Studies, 87(5), 2020

[6] *OTMedia, l'observatoire transmédia de l'actualité*, N. Hervé, Culture et Recherche n°141, 2020

[7] Comment Didier Raoult et la chloroquine ont surgi dans le traitement médiatique du coronavirus, A. Bayet, N. Hervé, La Revue des Médias, 2020

[8] *Représentations lexicales pour la détection non supervisée d'événements dans un flux de tweets : étude sur des corpus français et anglais*, B. Mazoyer, N. Hervé, C. Hudelot, J. Cagé, EGC 2020

[9] *Réduire les biais dans la collecte de tweets*, B. Mazoyer, N. Hervé, C. Hudelot, J. Cagé, EGC, journée DAHLIA 2019