Arcom

Réponse à la consultation publique sur l'accès aux données des plateformes pour la recherche

Agence France Presse

A.1. Avez-vous déjà mené des recherches utilisant des données issues d'une ou plusieurs plateformes en ligne ? Si oui, comment les avez-vous collectées (par exemple à l'aide d'API, de crowdsourcing, etc.) ?

Oui. Via CrowdTangle pour Facebook, Instagram, Reddit, via scraping ou APIs pour Twitter, via API pour YouTube.

A.2. Avez-vous rencontré des difficultés dans la collecte de ces données ? Si oui, de quel ordre ? Donnez des exemples.

Il y a des limitations inhérentes à chaque plateforme, que ce soit la temporalité d'accès aux données, les mesures d'audience figées au jour et heure de la requête, la possibilité ou pas de rechercher des urls ou des noms de domaine sur un laps de temps donné.

Facebook et Instagram prennent beaucoup de mesures (changement de code) pour éviter le scraping (y compris des avertissements en rouge dans le code de leurs pages).

A.3. Si oui, avez-vous déjà abandonné tout ou partie d'un projet de recherche du fait de l'impossibilité d'accéder à des données de plateformes en ligne ? Si oui, était-ce la conséquence d'un refus d'accès ? Donnez des exemples.

Nous avons abandonné une tentative de quantifier les principaux partages de liens (urls) sur Facebook pendant la campagne électorale française, faute de pouvoir automatiser la collecte via l'API de CrowdTangle. Même en ayant accès à l'API, la version pour les médias est semble-t-il plus bridée que certains accès utilisés par des universités.

A.4. Si non, quels ont été selon vous les facteurs qui vous ont permis de collecter ces données de manière fructueuse? Avez-vous pu bénéficier de la collaboration de la plateforme étudiée pour accéder à ces données ? Si oui, comment s'est-elle matérialisée ? Donnez des exemples.

A.5. Avez-vous établi une politique de partage de vos données avec des tiers à des fins de recherche ?

i) Si oui :

- *- depuis quand existe-t-elle ?*
- *- concerne-t-elle une ou plusieurs catégories de bénéficiaires particuliers (chercheurs, ONGs, entreprises, etc.) ?*
- *- existe-il des critères de sélection de ces bénéficiaires ? Si oui, lesquels ?*

- quel(s) type(s) de données cette politique concerne-t-elle ?
- intègre-t-elle un volet de contrôle ou de suivi de l'utilisation qui est

faite des données délivrées ?

ii) Si non, quelles sont les raisons pour lesquelles vous n'avez pas initié une telle politique ? Il peut notamment s'agir de risques d'ordre juridique, réglementaire, technique, financier, etc. Précisez quelle a été votre évaluation de ces risques menant à la décision de ne pas ouvrir vos données.

Nous travaillons actuellement dans le projet DeFacto à l'élaboration de bases de données partagées où nous essayons d'augmenter / enrichir ces données à l'aide de celles dont disposent les vérificateurs (fact checkeurs) mais de façon à ne pas dévoiler la provenance de ces rajouts afin d'éviter tout risque juridique.

B.1. Doit-on définir et éventuellement limiter en amont les types d'acteurs pouvant bénéficier d'un accès à des données : chercheurs, journalistes, ONGs, think tanks, société civile, etc. ?

i) Si oui, selon quels critères (éventuellement combinés à la nature même de la recherche ou des objectifs poursuivis) ?

Tout média, organisation de fact-checking, ONG ou chercheur travaillant sur la désinformation devrait pouvoir accès aux données, en se constituant leurs propres jeux de données sur la base d'une recherche de leur intérêt, sans autorisation préalable ou procédure bureaucratique.

ii*) Doivent-ils avoir les mêmes possibilités d'accès ou bien celles-ci doivent-elles différer selon le type d'acteur ?*

Les mêmes possibilités d'accès.

B.2. Doit-on également définir un niveau minimal d'accès à destination du grand public (ou d'une catégorie de bénéficiaires plus large que le champ strict des chercheurs académiques) telle que la mise à disposition obligatoire d'un certain nombre de données anonymisées en open data ?

La difficulté est de savoir si ces données même anonymisées pourraient être utilisées aussi par des acteurs / auteurs de désinformation pour mieux cibler leurs campagnes, améliorer leur amplification, anticiper les contre-narrations (fact-checking) etc.

Par contre, le grand public devrait avoir un niveau minimal d'accès aux contenus repérés comme de la désinformation, du discours de haine ou hyperpartisan. L'archivage de ces contenus, en tant que preuve de la manipulation de l'information, est l'un des gros problèmes actuel des fact checkeurs parce que plusieurs plateformes (notamment Facebook et Instagram) mettent très fréquemment leur code à jour ce qui a pour effet de limiter l'archivage (et la preuve de la manipulation).

B.3. Selon vous, un modèle d'accès à des données via la formulation des demandes d'accès à un tiers de confiance est-il pertinent ?

i)

-

Si oui :

ce tiers de confiance devrait-il être un acteur public européen ou national ? Dans ce cas, quelles seraient ses interactions avec les autres autorités, par exemple celle(s) en charge de la protection des données personnelles ?

quelles pourraient être les modalités d'organisation d'un protocole fléché et encadré d'accès aux données ?

-

Les modalités d'implication du tiers de confiance seraient-elles à définir selon le niveau de risque associé aux données ?

ii) Si non :

- *- pour quelles raisons ? Celles-ci peuvent être diverses : juridique, académique, logistique, etc.*

non, parce que trop bureaucratique et trop lent pour répondre à la désinformation. Un contenu peut devenir viral parfois plusieurs jours après sa première diffusion, le temps de trouver des relais suffisamment puissants, de profiter d'une conjoncture favorable (en lien avec l'actualité par ex.). On ne connaît le niveau de risque qu'après analyse et encore… On ne peut pas prévoir à l'avance ce qu'on va trouver dans les données : il y a une découverte empirique, qui est du ressort de la sérendipité et de l'analyse a posteriori.

- *- un modèle d'interaction direct entre la plateforme et les chercheurs vous apparaît-il préférable ? Si oui, pourquoi ?*

  C'est préférable parce que plus direct et cela permet d'être plus réactif. Il est important que les chercheurs puissent fabriquer / extraire eux-mêmes leurs jeux de données en fonction de leurs objectifs mais aussi en fonction de ce qu'ils ou elles découvrent dans les jeux de données, pour approfondir. (par ex, on peut à partir d'une première requête, découvrir qu'un autre hashtag co-référent dans le même corpus est beaucoup plus viral et ramène davantage de données).

- *B.4. Dans l'hypothèse d'un mode de régulation qui impliquerait l'intervention d'un tiers de confiance dans l'ouverture des données pour des projets de recherche :*

1. *i) qui aurait la charge de contrôler la mise en œuvre du protocole de demande ?*

   Il devrait y avoir un suivi de la part du tiers de confiance et un retour au fil de l'eau de la part des projets de recherche pour savoir s'ils ou elles pensent avoir obtenu satisfaction.

2. *ii) quels garde-fous pourraient être mis en place pour assurer un accès à des données permettant de répondre au besoin exprimé de manière satisfaisante ?*

   Un suivi régulier et des sanctions administratives contre les plateformes (d'un montant équivalent au coût déclaré du projet) pour celles qui ne fourniraient pas les données nécessaires à la réalisation du projet de recherche (dans la mesure où ces données sont disponibles et peuvent être extraites).

3. *iii) comment la transparence des décisions des organisateurs du protocole d'accès devrait-elle être garantie ?*

   Les organisateurs devraient s'assurer, lors de la soumission d'un projet de recherche nécessitant des données des plateformes numériques, que les demandes sont clairement exprimées et que les plateformes sont a priori en mesure de les satisfaire.

4. *iv) quelle place et quels rôles devraient avoir chacune des parties prenantes et notamment les plateformes ?*

   Le tiers de confiance et les plateformes devraient pouvoir valider l'accès aux données pour éviter qu'un projet de recherche ne se construise sur un accès hypothétique à des données qui ne seraient pas ensuite disponibles (ou prises en compte dans les interfaces et/ou APIs mises en place).

   Les organisateurs devraient pouvoir contre-argumenter auprès des plateformes si celles-ci ne jouent pas le jeu.

5. *v) identifiez-vous des risques inhérents à ce modèle ? Lesquels ?*

Une étape bureaucratique supplémentaire qui risque de ralentir les travaux de recherche. Un effet « ardoise magique » : bon nombre de contenus litigieux, vérifiés comme faux par les fact checkeurs, sont souvent effacés par leurs auteurs, voire par les plateformes pour des raisons du non respect de leurs conditions d'utilisation. Des pans entiers de désinformation pourraient ainsi disparaître subrepticement des corpus mis à disposition.

C.1.Lors de l'élaboration de leur(s) demande(s) d'accès, comment favoriser la connaissance des chercheurs des données des plateformes qu'ils pourraient solliciter pour leurs études ?

C'est l'intérêt d'avoir une interface de recherche (de type CrowdTangle) permettant aux chercheurs de faire des requêtes et de récupérer des corpus par eux-mêmes, au besoin après une formation leur indiquant ce qu'ils peuvent récupérer et comment utiliser au mieux ces outils. L'amélioration régulière de ces interfaces, suite au retour des chercheurs devrait également être envisagée.

C.2. Qui définirait le contour des projets de recherche et leur rattachement à une ou plusieurs missions d'intérêt général et présidant à l'identification des données auquel l'accès serait nécessaire? Les données concernées doivent-elles être restreintes à des champs de recherche particuliers ? Si oui, lesquels ? Par exemple, lutte contre la manipulation de l'information, la haine et le piratage en ligne.

Le tiers de confiance (Arcom?) pourrait maintenir une liste des chercheurs, journalistes, fact checkeurs, ONGs de la société civile, habilités à mener de telles recherches dans la lutte contre la désinformation, contre le discours de haine, le complotisme, etc. auxquels les plateformes devraient ouvrir accès à leurs interfaces ou APIs sans rechigner et sans délai.

Comme Facebook utilise la certification de l'IFCN (signataires du Code of principles) pour recruter les membres de son réseau de tiers vérificateurs, l'Arcom (ou autre tiers de confiance) pourrait pré-sélectionner les chercheurs, journalistes, organisations (…) habilités à accéder aux données des plateformes.

C.3.Comment seraient formulées les demandes d'accès par les chercheurs intéressés ? Par exemple via des appels à projets sur des thématiques prédéfinies et/ou ad hoc, après identification de sujets d'étude pertinents ?

De nouvelles demandes d'accès pourraient être formulées via des projets de recherche si les chercheurs impliqués ne sont pas déjà répertoriés.

C.4. Jugez-vous pertinent l'intervention d'un comité d'évaluation et de suivi des demandes d'accès ?

i) Si oui, comment devrait être composé ce comité d'évaluation (par exemple un comité scientifique international) ? Un ou plusieurs régulateurs devraient-il y avoir une place et un rôle et, si oui, lequel ?

ii) Si non, pourquoi ? Quels dispositifs vous sembleraient plus à même de répondre aux demandes d'accès des chercheurs ?

Encore une fois, bon nombre de contenus repérés comme faux ou litigieux auront disparu entretemps (ou auront migré vers d'autres plateformes). Le meilleur dispositif serait une liste de chercheurs accrédités / habilités maintenue par le tiers de confiance.

C.5. Dans quelle mesure le caractère plus ou moins contraignant des obligations d'ouverture de leurs données pour les plateformes impliquerait-il leur présence dans les comités d'évaluation ? Faut-il également que les plateformes bénéficient d'un droit de retour par rapport aux demandes des chercheurs voire d'un droit de refus ?

L'extraction de données présuppose de respecter la réglementation en vigueur et notamment le RGPD mais également une faisabilité technique, par ex. si les données demandées ne sont pas collectées ou seulement partiellement ou si leur complexité peut être source d'erreurs. Les plateformes pourraient exercer un droit de retour lors de la présentation préliminaire des travaux de recherche devant un comité scientifique (et éthique ; voir C8, C9).

C.6. Quels seraient les critères d'attribution des accès ? Par exemple, obligation pour être sélectionné d'avoir un projet de recherche mobilisant des équipes interdisciplinaires, éventuellement issues de structures implantées dans au moins deux pays de l'Union européenne ?

Tout critère bureaucratique de ce genre risque de faire le jeu des plateformes qui, en définitive, traitent le problème de la désinformation comme un accident de parcours et de relations publiques. Tout ce qui entrave la mise à disposition des données pour des chercheurs indépendants va dans le sens d'un effet « ardoise magique » au bénéfice des plateformes.

Soit dit en passant, concernant les critères d'attribution, un projet européen comme l'observatoire DeFacto, est composé uniquement de partenaires français.

C.7. Faut-il inclure une dimension temporelle dans l'évaluation des appels à projets pour ne retenir que ceux au temps court ou long ?

Il faut une dimension temporelle et la plus longue possible, notamment pour être en mesure de remonter, si possible, à l'origine des campagne de désinformation. Toutefois, la difficulté de limiter l'accès aux données aux seuls projets de recherche, c'est donner la clef aux plateformes pour valider la mise à disposition des données qu'elles veulent bien (ou pas) partager.

C.8. Les travaux issus de l'analyse de ces données doivent-ils bénéficier d'une certification externe ? Si oui, quelle forme pourrait-elle prendre ?

Il y a un principe scientifique de revue par les pairs. Un comité scientifique (et éthique) révisant les travaux issues des données collectées via la certification émise par le tiers de confiance pourrait être envisagé.

C.9. Quelles doivent être les précautions à prendre en ce qui concerne la publication des études menées, par exemple eu égard à la sensibilité des données qui auraient été exploitées ? Comment conjuguer la mise en œuvre de ces mesures de précaution et le principe fondamental d'indépendance des chercheurs ?

Le comité scientifique évoqué en C8 pourrait émettre des recommandations sur les précautions à prendre tant au niveau de la collecte, de l'analyse et de l'interprétation.

D.1. Compte tenu du fait que les projets de recherche s'appuyant sur des données de plateformes peuvent privilégier un angle d'analyse qui rendrait nécessaire un format spécifique des bases des données (variables, granularité, etc.) :

1. *i) comment permettre la création de bases de données spécifiques ou uniques qui seraient construites pour répondre à des besoins précis ?*

Le fonctionnement actuel de CrowdTangle permet d'extraire des fichiers csv (limités à 10.000 lignes et des périodes temporelles plus ou moins réduites selon les requêtes) que les chercheurs peuvent ensuite utiliser, enrichir, compléter en fonction de leurs besoins (et de la disponibilité des annotations, données complémentaires …).

C'est une base intéressante même s'il y a trop de limites volumétriques (limitation à 10.000 lignes) et temporelles (par ex. sur l'analyse de noms de domaine de sites de désinformation pour analyser leur propagation dans le temps par ex.)

2. *ii) dans quelle mesure certains projets de recherche permettraient-ils de construire des indicateurs ou mesures innovants qui pourraient participer à la connaissance collective des problématiques étudiées ?*

Construire des indicateurs ou mesures innovants présuppose un accès rapide et sans restriction aux données ainsi que la possibilité de faire un suivi sur un laps de temps donné (par ex. une élection), de préférence de façon programmatique (afin d'éviter les collectes manuelles fastidieuses sur de longues périodes de temps).

D.2. Les accès aux données peuvent-ils et doivent-ils faire l'objet d'une co-construction à part égale entre acteurs de la gouvernance, chercheurs et plateformes sur le modèle du CASD de l'Insee[15] ?

C'est à chaque plateforme de définir les modalités d'authentification et d'accès aux données et à la gouvernance de donner aux plateformes la liste des personnes habilitées à y accéder ainsi que la liste des données à mettre à disposition. Il s'agit, non pas de statistiques collectées sur des individus, mais de billets publiés en général publiquement par leurs auteurs.

Se pose en revanche le problème des groupes ou pages Facebook ou Whatsapp qui, eux, échappent en bonne partie à l'ouverture des données.

D.3. Comment le cadre d'accès aux données – gouvernance, types de données identifiées en lien avec les missions, etc. – peut-il être rendu pérenne afin de rester adapté aux innovations et évolutions régulières des plateformes ?

On pourrait définir comme cadre que tout ce qui est visible par n'importe quelle personne dans son navigateur (au besoin avec login comme sur FB) est une donnée publiquement accessible et qu'elle doit l'être aussi dans une interface de requêtes destinée aux chercheurs, sans restriction.

Par ailleurs, concernant le type de données (seule question sur ce point important?), il est essentiel que les chercheurs, journalistes, fact checkeurs, aient aussi accès à des données sur les images et vidéos (y compris les mèmes internet) sur leur propagation, leur amplification, leur reproduction, leurs statistiques évolutives de partage (…). La plupart des plateformes ont des index de similarité d'images mais les gardent en interne (FB, Instagram) tandis que l'indexation des contenus par plateforme par des moteurs tels que Google Images, Google Lens, Bing, Yandex, Tineye, est peu documentée voire très parcellaire pour les agoras les plus fermées.

D.4. Quels modes d'accès devraient être privilégiés pour les données de plateformes en ligne ? Quels sont leurs différents avantages et inconvénients ? Ceux-ci doivent-ils différer selon les données collectées ? Si oui pourquoi ?

Interface de recherche et collecte ainsi qu'une (ou plusieurs) API pour la collecte programmatique. Chaque plateforme aura ses propres interfaces ou APIs.

D.5. Comment garantir un mode d'accès sécurisé, notamment lorsque les données ne sont pas anonymisées et/ou touchent à des problématiques de secret des affaires ?

Tout ce qui est visible publiquement dans un navigateur (connecté à la plateforme ou pas), devrait être ensuite visible et récupérable pour les chercheurs (voir sur ce point E4 1) i) plus bas).

D.6. De quelle manière devraient être stockées ces données afin d'assurer la protection des données personnelles et, le cas échant, du secret des affaires ?

Huma-Num, en gérant les permissions, les mises à disposition publiques de données anonymisées si besoin.

D.7. Quel serait le rôle et le champ d'intervention des autorités de protection des données (nationales et du Centre Européen de Protection des Données) dans l'évaluation des risques associés à l'accès à ces données ?

Si cela rentre dans leur champ de compétences, elles pourraient être consultées si un projet de recherche voulait (re)publier des données précédemment collectées auprès d'une ou plusieurs plateformes dans ce cadre.

D.8. Les projets de recherche doivent-ils bénéficier d'un soutien de la part de la structure qui serait en charge de l'attribution des accès, par exemple de nature technique, financière ou autre ?

Si une telle structure était créée, elle pourrait conseiller les chercheurs notamment sur le plan technique voire épistémique sur les meilleurs pratiques de récupération des données convoitées.

E.1. Comment accompagner les chercheurs dans la construction de leurs projets de recherche et leur mise en conformité avec le RGPD et les normes établies par le dispositif ?

Si la structure de D.8 était créée, elle pourrait réaliser cet accompagnement et mutualiser la connaissance globale sur l'accès aux données pour la recherche.

E.2. Quels dispositifs permettraient d'atténuer les écarts de financement et de capacité techniques entre institutions académiques pouvant déboucher sur une captation des projets par un nombre restreint d'équipes de recherche ?

Maintenir une liste de chercheurs, journalistes, fact checkeurs accrédités et donc habilités à faire des requêtes et à récupérer des données.

Multiplier les autorisations, les maillons intermédiaires, les comités, contribuera à éliminer bon nombre d'équipes de recherche, de chercheurs ou d'institutions et risque d'aboutir à des corpus de données amoindris et donc à des analyses partielles voire biaisées. On aboutirait à une moindre documentation de la désinformation, sauf à croiser les données des plateformes avec celles des fact-checkeurs. On risquerait de perdre une bonne partie de la mémoire de la désinformation, mémoire prise au sens de « système de projection d'une information dans le futur » (Stanislas Dehaene).

E.3. Comment mettre en place des incitations efficaces et équilibrées pour que les plateformes s'inscrivent dans des dynamiques d'ouverture des données ? Comment intégrer ces acteurs dans le dispositif de manière cohérente et favoriser les bonnes pratiques ?

Des représentants des plateformes devraient siéger dans le comité scientifique (et éthique) pour donner leur avis sur les recherches en cours ou finalisées, avec un rôle consultatif qui pourrait éviter des erreurs ou biais liés à une méconnaissance des données mises à disposition et pourrait les rassurer quant à l'ouverture des données.

E.4. L'intervention d'un comité d'audit externe serait-elle pertinente :

1. *i) en amont, dans le cadre de l'évaluation des décisions d'agrément sur le modèle par exemple du CESP dans le champ des enquêtes statistiques en France ?*

Le CESP traite de données biostatistiques collectées sur des patients. Dans le cas des plateformes, il s'agit de comprendre la propagation et les effets potentiels de messages comportant des données

potentiellement nocives (désinformation, discours de haine ou hyperpartisan) partagées publiquement (ou dans des groupes) sur une agora privatisée. Ces données sont « brutes » et non générées statistiquement. Elles sont librement publiées sur ces plateformes par les utilisateurs eux-mêmes. Pourquoi faudrait-il prendre autant de précautions ? Pourquoi un contenu diffusé publiquement avec une intention propagandistique deviendrait subitement un contenu privé inaccessible, notamment après avoir démystifié ?

2. *ii) en aval, dans l'examen des réponses apportées par les plateformes aux demandes d'accès ?*

Les plateformes ne devraient pas jouer un rôle d'arbitre dans l'accès aux données. Cet accès devrait être libre pour les chercheurs, journalistes, fact checkeurs, accrédités et travaillant sur la désinformation en toute indépendance.

E.5. Quelles garanties procédurales pourraient être mises en place en lien avec les problématiques de secret des affaires ?

Réponse de non-juriste : révéler que de la désinformation, des discours de haine ou hyperpartisans circulent, se propagent, sur des réseaux sociaux à cause de biais algorithmiques, de phénomènes d'amplification, ne devraient pas pour autant constituer une atteinte au secret des affaires dans la mesure où il n'y a pas de divulgation de documents internes, d'informations à valeur commerciale, d'informations ayant fait l'objet de dispositions pour les garder secrètes, ou d'appropriation d'un savoir-faire (loi n°2018-670). Cela relève plutôt de la critique légitime d'un produit ou un service ne donnant pas les résultats attendus.