Arcom

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

Medialab – Sciences Po

Réponse médialab à la consultation lancée par l'ARCOM

Préambule :

Le rapport suivant a été produit par un groupe de chercheur.e.s au médialab en juin 2022 : https://hal-sciencespo.archives-ouvertes.fr/hal-03711842. Il a pour objectif de discuter des problématiques autour des actions des grandes plateformes en matière de désinformation. L'Annexe 1 (p.18) du rapport fait la synthèse des données et des informations manquantes pour pouvoir étudier et mener des tâches d'audit concernant les différents types d'interventions ayant pour but de modérer les contenus. Ainsi, le rapport couvre une partie des questions posées par la consultation dans les sections A (partage d'expérience d'utilisations de données) et la section B (gouvernance).

A Partage d'expériences d'utilisations de données des services en relation avec la thématique

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.) ?

Le médialab réalise de très nombreuses collectes de données en ligne visant à alimenter différents travaux de recherche. Du scraping à l'utilisation d'APIs dédiées ou non, la plupart des possibilités technologiques ont déjà pu être mises en œuvre en respectant les fonctionnalités et les modalités prévues par chaque plateforme. Le médialab publie à ce titre plusieurs outils logiciels en Open Source permettant de simplifier et automatiser ces tâches (citons notamment "minet" et "gazouilloire", conçus pour réaliser simplement des collectes ponctuelles ou massives de données sur de nombreuses plateformes et notamment Twitter).

À titre d'exemple :

- Facebook : utilisation des APIs fournies par l'intermédiaire de CrowdTangle, exploitation des données mises à disposition de certains académiques via la base de données Condor par le projet Social Science One, et scraping de commentaires depuis des pages et groupes Facebook accessibles publiquement
- Twitter : utilisation des différentes API gratuites (v1.1 et v2) et réservées aux académiques, mais également scraping exploitant l'API interne du site web, publiquement accessible
- YouTube : exploitation des APIs gratuites pour l'accès aux métadonnées et commentaires des vidéos et chaînes, et scraping pour l'accès aux sous-titres des vidéos
- TikTok : scraping exploitant l'API interne du site web, publiquement accessible

Contrairement à de nombreux laboratoires anglo-saxons, notre usage est resté depuis toujours exclusivement gratuit : le laboratoire n'a jamais payé pour bénéficier d'accès spécifiques discrétionnaires ou plus volumineux que ne le permettent les accès publics.

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

Oui, notamment :

- problèmes de limitation du volume de données qu'il est possible de récupérer (souvent à cause d'un *rate limiting* très agressif) ;
- problèmes d'APIs mal conçues, ou du moins pas conçues pour un usage de recherche (par exemple, la plateforme CrowdTangle est conçue pour des usages de *social listening* et n'est pas adaptée à la collecte transversale de données quand on souhaite par exemple collecter des posts sur plusieurs années et pas seulement les derniers posts en tendance) ;
- problèmes d'APIs donnant accès à des échantillons potentiellement non représentatifs de l'ensemble de ce dont dispose la plateforme (CrowdTangle, par exemple, ne fournit des informations que sur les pages et groupes Facebook les plus importants ou ceux spécifiquement suivis par leurs utilisateurs) ;
- problèmes d'APIs conçues de manière hostile (souvent pour empêcher la concurrence de récupérer les données) en interdisant par exemple toute énumération efficace (routes de recherche ne permettant pas de délimiter temporellement les requêtes ou avec une limite très basse du nombre maximum de résultats renvoyés) ;
- problème d'accès à certaines métadonnées non proposées dans les APIs (métadonnées que l'on peut des fois, ironiquement, trouver en accès libre sur le site web et qu'il est donc possible de collecter par scraping) ;
- problème d'accès aux paradonnées (i.e. données auxiliaires sur les processus de création, enrichissement et destruction des données) relatives par exemple aux données disparues (motifs d'attrition, censure ou autosuppression, etc.), aux informations sur le fonctionnement des algorithmes de recommandation et l'ordre de présentation des résultats (par exemple YouTube ou TikTok), ou encore aux données d'audience (des plateformes, des agrégats sur les plateformes tel que les comptes, groupes, etc., ou même des posts individuels) : on dispose souvent d'un nombre de likes, shares ou de commentaires, mais rarement de vues ou clics effectifs ;
- restrictions techniques imposées par certains acteurs pour accéder à leurs données, complexifiant considérablement l'exploitation (salles physiques non connectées à Internet, accès via un VPN à des ressources serveur insuffisantes et limitant les technologies qu'il est possible d'employer pour réaliser les calculs, impossibilité de rapatrier le résultat des calculs, etc.).

A contrario, la plateforme Wikipédia offre un accès ouvert à l'ensemble de ses données, ce qui a constitué un terrain très favorable au développement des *Wikipedia Studies*.

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.

Les projets de recherche sont plutôt conçus en fonction des données déjà identifiées comme disponibles ou atteignables et en fonction des efforts techniques requis pour en effectuer la collection. Cela a tendance à créer un biais d'opportunité qui explique par exemple que de nombreux projets préfèrent travailler sur des données Twitter plutôt que sur Instagram, Whatsapp, Snapchat, etc. Des projets ne sont donc pas abandonnés du fait d'un refus d'accès aux données, mais ceux-ci ne sont jamais formulés. De même, du fait de l'asymétrie de la relation entre recherche et plateformes, il est extrêmement rare d'envisager demander des données à une plateforme car il est souvent évident que la réponse sera négative.

Dans un ordre d'idée similaire, il est arrivé à certains de nos collègues de devoir abandonner un projet de recherche en partenariat avec une plateforme, une fois révélées des erreurs irrécupérables dans les données mises à disposition. Ceci est arrivé par exemple à nos collègues américains travaillant sur les bases de données Condor de Facebook, mises à disposition via le projet *Social Science One*. Il fut en effet annoncé, plus de deux ans après la mise en route du projet, que les données d'audience américaines étaient fortement biaisées et lacunaires à cause d'une erreur de calcul à la création de la base, rendant ainsi caduques les résultats de recherche s'y rapportant.

Notre expérience du dispositif Social Science One nous conduit par ailleurs à souligner le décalage considérable entre les intentions initiales du projet (permettre aux chercheurs de travailler dans des conditions sécurisées sur un ensemble de données de Facebook a granularité forte) et le résultat final. Les données auxquelles accèdent les chercheurs ont été si fortement anonymisées et agrégées qu'elles ne permettent pas de remplir la promesse initiale. Cette agrégation des données (sur un nombre restreint de variables) ne permet pas de conduire des recherches approfondies. Les données collectées à travers Crowd Tangle présentent souvent un intérêt bien plus grand pour la recherche - et Facebook est cependant en train de fermer cet accès important aux données.

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.

Dans quelques cas, le médialab a pu bénéficier d'accès privilégiés à certaines données de plateformes spécifiques sous la forme d'accords discrétionnaires.

Comme partie prenante du dispositif Social Science One, le médialab a accès à la base de données Condor qui fournit des données d'audience relatives au partage de certaines urls sur Facebook (nombre de vues, de likes, de clics, envoi à des *fact-checkers*, signalement par des utilisateurs, etc.).

De même, lors d'un évènement en partenariat avec Google France, une clé d'API avec volumes élargis avait été fournie afin de collecter des données YouTube dans des volumes supérieurs à ceux classiquement accessibles gratuitement, mais là encore le scraping s'est ironiquement avéré plus efficace.

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 ?

Adepte de l'*Open Science,* le médialab cherche autant que possible à assurer la reproductibilité de ses travaux, notamment en republiant en *Open Data* autant que possible les données sur lesquelles s'appuient ses travaux. Ainsi, les corpus de données de recherche du médialab sont notamment mis à disposition de la communauté de recherche sur l'entrepôt institutionnel dédié data.sciencespo.fr.

Cependant, les données issues des plateformes imposent souvent un certain nombre de réserves afin de pouvoir rediffuser les données, ce qui contraint les conditions légales de repartage : questions de copyright, enjeux de vie privée et de respect du RGPD dès lors que les données sont rattachables à des individus (ré)identifiables, clauses spécifiques des CGU, etc.

À titre d'exemple, les CGU de l'API de Twitter interdisent de rediffuser publiquement autre chose que les identifiants uniques des tweets considérés au sein d'une étude. Cela pose un premier problème concernant la reproductibilité car la forte attrition des messages postés sur Twitter (on estime à 20% la proportion de tweets supprimés par leurs auteurs en moins d'un an) empêche alors de reconstituer les corpus de travail à l'identique. Plus ennuyant encore, ces CGU limitent la taille de ces corpus d'identifiants de tweets à 1,5M de tweets et une période de 30 jours, rendant de fait la publication quasi impossible dans le cas de très nombreux travaux.

Dans ces différents cas, seules les données légalement rediffusables sont redistribuées publiquement. Pour autant il peut arriver, au cas par cas, projet par projet, que des données plus complètes soient repartagées à des tiers, mais uniquement dans le cadre de partenariats de recherche spécifiques clairement établis, et avec déclaration explicite de ces partenaires dans la déclaration RGPD associée.

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.

Certains jeux de données utilisés par les recherches du médialab ne peuvent pas être partagés en raison des CGU des plateformes (*supra*), de contraintes liées à la propriété intellectuelle (par exemple les textes des articles de presse collectés dans certaines recherches). Ces contraintes nous semblent souvent déraisonnables et conduisent à limiter le partage de jeux de données, une pratique essentielle pour développer un écosystème de recherche large et diversifié, et nécessaire pour assurer la reproductibilité des résultats. Les revues scientifiques et les organismes de financement de la recherche - à juste titre - exigent de plus en plus l'ouverture des données de recherche.

Remarque sur le partage d'expérience

Il faut noter que les organismes de recherche, ONGs, journalistes, etc. ont toujours eu recours à des méthodologies de collecte de données se pratiquant à l'insu des plateformes elles-mêmes. Le scraping en est un exemple parfait et permet de collecter des données alors même que les plateformes ont souvent une attitude antagoniste à l'égard des acteurs mentionnés ci-avant. Cette pratique, souvent considérée à tort comme illégale ou peu éthique, est néanmoins nécessaire pour garantir une forme d'indépendance envers les plateformes elles-mêmes. Aussi avons-nous constaté ces dernières années une tentative accrue d'attaquer en justice des chercheurs utilisant ce genre de méthodes dans des procès (aux Etats-Unis

surtout, notamment celui perdu par LinkedIn) tentant d'empêcher des entreprises légitimes de collectes de données pour des besoins de recherche et d'utilité publique (à noter qu'un pan de cette judiciarisation porte aussi sur des questions de concurrence entre entreprises accusées de se voler des données). Il serait donc utile que ces pratiques, notamment celle du scraping, soient définitivement protégées et pérennisées dans le cadre d'une exception de recherche empêchant aux plateformes d'attaquer des équipes de recherche indûment.

B Gouvernance

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. ?

Dans l'idéal, nous estimons que l'ensemble des données ni sensibles, ni personnelles, soit anonymisées, soit agrégées, et non soumises à des contraintes légales de type droit d'auteur, **ont vocation à être accessibles gratuitement et sans limite en Open Data par tous**.

Concernant les données sensibles, il est essentiel de restreindre les capacités et les conditions d'accès à certains types d'acteurs uniquement. Il est important pour un certain nombre de travaux de recherche, journalistiques, ou de contrôle associatif et citoyen, de pouvoir accéder à ces données, mais il est également important de s'assurer que ces statuts ne contribuent pas à donner un accès libre à n'importe quel acteur qui pourrait en détourner l'usage à d'autres fins (comme l'exemple de l'affaire Cambridge Analytica l'a bien illustré). Ce risque justifie à nos yeux l'existence d'une gouvernance spécifique de l'accès aux données qui ne sont pas ouvertes en Open Data par les plateformes.

L'accès aux données plus "sensibles" à des acteurs privés (notamment venant du Search Engine Optimization, SEO et Social Media Optimization, SMO) n'est pas une question de notre ressort. Les plateformes ouvrent déjà de nombreuses données aux acteurs marchands pour leurs activités - ceux-ci ont parfois accès à des données d'une précision beaucoup plus grande que celles dont disposent les chercheurs. Il est probable que donner un accès renforcé aux acteurs commerciaux augmenterait les possibilités de "jouer" avec les plateformes au profit de leurs clients. Il nous semble que l'accès aux données des plateformes dans le cadre du DSA ne devrait pas être autorisé pour les acteurs qui les utiliseraient pour mener des actions (marketing, publicité, stratégie éditoriale) à l'égard des utilisateurs.

Par ailleurs, toute cette discussion doit être associée à une typologie du niveau de risque de la diffusion de ces données. Au médialab, cette question a fait l'objet d'une réflexion conduite par Shaden Shabayek, Emmanuel Vincent et Héloïse Théro à la suite de plusieurs recherches portant sur l'audit du comportement de Facebook lorsque des informations fausses ont été signalées à la plateforme par des fact-checkers. Ces recherches ont fait l'objet de publications :

- *Théro, H. & Vincent, E. M. (2002). Investigating facebook's interventions against accounts that repeatedly share misinformation. Information Processing and Management, 59(2).*
- *Vincent, E. M., Théro, H. & Shabayek, S. (2022). Measuring the effect of Facebook's downranking interventions against groups and websites that*

repeatedly share misinformation. Harvard Kennedy School (HKS) Misinformation Review, 3(3).

Un policy report indiquant le type de données auxquelles les chercheurs ne peuvent accéder alors qu'elles seraient nécessaires pour conduire une analyse rigoureuse des plateformes (Facebook, YouTube, Twitter) a été publié ainsi qu'un article de recherche plus complet :

- *Shaden Shabayek, Héloïse Théro, Dana Almanla, Emmanuel Vincent. Monitoring misinformation related interventions by Facebook, Twitter and YouTube: methods and illustration. 2022. ⟨hal-03662191⟩*

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

Il nous semblerait plus pertinent de ne pas définir l'accès aux données par le seul statut des acteurs demandant un accès aux données mais de le faire à partir de projets de recherche déposés devant une instance d'évaluation. Celle-ci doit pouvoir vérifier que le projet a bien un but de production de connaissance d'intérêt public. La base légale de l'utilisation des données par les chercheurs (ou les journalistes et les ONG) doit être celle de l'exécution d'une mission d'intérêt public, telle que définie par le RGPD : une production de connaissance qui « n'est pas utilisée pour prendre des décisions à l'égard des personnes concernées » (article 4.2° de la loi Informatique et Libertés).

Pour autant prévoir un dispositif simplifié d'accès privilégié récurrent et stable pour les équipes de recherche spécialisées dans le domaine pourrait évidemment être intéressant.

Dans tous les cas, il est indispensable que cette évaluation des projets et équipes de recherche soit réalisée de la manière la plus indépendante, transparente et pluraliste possible par un comité représentatif large.

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

Il est probable que certains projets, en raison de l'accès à des données personnelles, soient mieux couverts par les obligations professionnelles des chartes de déontologie de la recherche et des codes de conduite. Seules les équipes de recherche disposent des compétences et d'un cadre déontologique permettant d'accéder aux données plus "sensibles" et granulaires des plateformes. Une solution pour les journalistes et les ONG serait éventuellement, pour ce type particulier de données, de travailler dans le cadre d'un partenariat piloté par un laboratoire de recherche - ce qui arrive dans le cas de projet ANR ou européens.

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* ?

Oui, comme indiqué en préambule à la question B.1.

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 ?

Oui, le dispositif qui nous semble le plus pertinent est de séparer deux instances :

- un comité d'évaluation des projets composé majoritairement de représentants des utilisateurs : académiques, associatifs et journalistiques, et de quelques représentants des régulateurs ;
- une instance définissant les conditions et les règles d'accès aux données. Deux configurations différentes peuvent être imaginées : (1) composée uniquement des plateformes et des régulateurs afin d'examiner les propositions sélectionnées dans le comité d'évaluation et de répondre des décisions d'accès devant lui ; (2) une composition associant plateformes, régulateurs et représentants des utilisateurs (chercheurs, ONG) afin que ces derniers puissent défendre les propositions du comité d'évaluation. Dans tous les cas, les décisions de cette instance doivent être présentées et justifiées devant le Comité d'évaluation.

Cette distinction permettrait de différencier l'évaluation scientifique des projets de celle des conditions d'accès aux données. Dans cette situation, les plateformes ne pourraient être présentes que dans le second comité et pas dans le premier.

Les projets proposés au Comité d'évaluation ainsi que les décisions prises par les deux instances devraient être rendus publics. Cela permettra la construction au fur et à mesure d'un registre public des données déjà disponibles et des cadres dans lesquels elles peuvent être mises à disposition des différents acteurs - au cas où un tel registre ne serait pas fourni d'emblée par les plateformes. Il serait utile également de considérer la mise en place d'une plateforme de suivi des projets avec les documents de travail produits. Cette plateforme pourra également servir de forum d'échange sur les problèmes rencontrés, par exemple sur la structure des données, la granularité des données ou encore le type de données auxiliaires qui ont pu être fournies, afin de formuler des demandes d'accès aux données de plus en plus pertinentes.

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 ?

Il semblerait plus pertinent, à la fois pour pouvoir peser auprès des plateformes, mais également pour tenir compte des nombreux projets collaboratifs internationaux, qu'un tel comité soit établi à l'échelon européen, tout en bénéficiant de relais et correspondants nationaux.

ii) Si non :

- pour **quelles raisons** ? Celles-ci peuvent être diverses : juridique, académique, logistique, etc.
- un modèle **d'interaction direct** entre la plateforme et les chercheurs vous apparaît-il préférable ? Si oui, pourquoi ?

Le modèle proposé plus haut présente le défaut de générer des processus qui peuvent être longs, fastidieux et handicapants pour la recherche. Il nous semble qu'il est cependant nécessaire pour peser de façon efficace sur les plateformes. En

revanche, un processus d'Open Data ouvert et sans gouvernance spécifique permettrait de faciliter l'accès aux données agrégées ou à faible niveau de risque.

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 :

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

Il semble préférable de donner la charge de faire appliquer les demandes validées par le comité d'évaluation par une instance de régulation distincte, disposant de pouvoirs de sanction judiciaire et/ou financière, à l'image des sanctions prononcées par exemple par la CNIL. Le montant de ces sanctions financières pourrait être systématiquement réaffecté au développement de la recherche publique.

i) 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 travail annuel de bilan et de contrôle de la bonne application des décisions passées avec retours utilisateurs est indispensable pour permettre au dispositif de rendre compte des difficultés rencontrées et pouvoir proposer des solutions d'amélioration.

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

Afin de garantir un maximum de confiance en l'institution, il semble essentiel que l'ensemble des projets soumis au comité, ainsi que ses décisions et délibérations soient rendus publics au plus tôt en toute transparence, par exemple sur le site du comité, mais également sous la forme de données open data réexploitables, versées dans une archive ouverte, permettant à la société civile de contrôler également l'activité du comité.

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

Cf proposition formulée en préambule de la question B.3.

iv) identifiez-vous des **risques inhérents** à ce modèle ? Lesquels ?

- Manque de moyens humains et financiers attribués au tiers de confiance ne lui permettant pas d'effectuer son travail correctement. Cela vaut aussi si l'on imagine le tiers de confiance comme un rassemblement d'acteurs concernés (recherche, ONGs, etc.) qui doivent avoir les moyens de consacrer du temps à ce travail administratif trop rarement pris en compte et qui s'impose souvent comme une charge de travail supplémentaire.
- Lourdeur des processus administratifs décourageant quiconque de formuler des demandes d'accès aux données.

- Faire attention aux conflits d'intérêt des ex-employés des plateformes qui pourraient être recrutés dans les structures du tiers de confiance, comme cela a pu être le cas dans le cadre du projet *Social Science One*, par exemple.

C. Construction des projets scientifiques

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 ?

Les plateformes devraient établir un **registre public des données** mises à disposition et de leurs différentes conditions d'accès.

Il est important d'instaurer un espace de discussion régulier avec les plateformes pour que les chercheurs puissent solliciter la communication de données nouvelles qui ne sont pas mises à disposition (demandes qui pourraient être adressées à l'intermédiaire en charge de la gouvernance de l'accès aux données - niveau européen). Il est très difficile pour les chercheurs de définir *a priori* la liste précise des données utiles à leur recherche ; très fréquemment, c'est en découvrant la base de données disponibles que des variables pertinentes pour la recherche apparaissent sans avoir été préalablement anticipées (cf. point A.3.).

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.

C'est le rôle du comité d'évaluation de juger de la pertinence de l'accès à certaines données en fonction des projets. Il ne serait pas souhaitable de limiter *a priori* les domaines ou sujets de recherche concernés. Les chercheurs eux-mêmes peuvent seuls en décider, quel que soit leur compétence ou discipline. À titre d'exemple, il est nécessaire d'avoir des données larges sur d'autres domaines pour mesurer la prévalence de phénomènes restreints comme la haine ou la désinformation.

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 ?

Ad hoc après identification de sujets d'étude pertinents. En effet, les appels à projets sur des thématiques prédéfinies impliqueraient de définir celles-ci au préalable et se poserait alors la question de qui doit les définir. De plus, des appels à projets imposeraient un calendrier qui pourrait venir en contradiction avec ceux des appels pour financement de projets et complexifier la réalisation de ces projets. Il semble plus sain de laisser les acteurs concernés (recherche, ONGs etc.) définir ce qui est pertinent et pouvoir solliciter les accès souhaités auprès du comité d'évaluation au fil de l'eau (à l'image par exemple des sollicitations de la CADA lors de demandes d'accès à des documents administratifs).

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 ?

Il est important de séparer :

- l'évaluation scientifique et l'intérêt général des projets (faite par un comité des pairs, et visant à évaluer l'objectif de produire de la connaissance) ;
- l'évaluation de la demande d'accès (faite par un comité pouvant comprendre à la fois les besoins, les régulateurs et les plateformes) ;
- le régulateur contraignant les plateformes à l'application des autorisations d'accès évaluées.

Il est par ailleurs important d'exclure les plateformes de l'évaluation scientifique des résultats (*peer review*, notamment), pour éviter qu'elles ne s'en servent comme un moyen possible de censure.

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

RAS

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** ?

Il est préférable que les plateformes ne participent pas au comité d'évaluation des projets, mais qu'elles soient en revanche présentes dans le comité validant l'accès aux données. Les résultats des projets devant être rendus publics, ils sont accessibles aux plateformes. Lorsque les projets mobilisent des données sensibles des plateformes, un aller-retour entre les équipes de recherche et les plateformes peut être mis en place afin de donner la primeur des résultats aux plateformes et leur permettre de formuler des remarques et d'éventuelles objections. Cette lecture prioritaire accordée aux plateformes peut leur permettre de formuler un droit de réponse mais ne peut leur permettre d'empêcher la publication des résultats.

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 ?

Les projets pluridisciplinaires et internationaux sont d'une grande importance et doivent bénéficier d'une attention spécifique. Cependant, en raison de la variété des sujets et de leur niveau différent de spécialisation, il est aussi crucial que des projets nationaux et mono-disciplinaires puissent être conduits.

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 ?

Non. La longueur d'un projet peut éventuellement peser dans l'évaluation du risque pour l'accès aux données, mais pas pour ne retenir qu'une temporalité de projets plutôt qu'une autre.

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 ?

Pour les travaux de recherche, la certification est apportée par la communauté (notamment à travers les publications et le *peer reviewing*). Il ne nous semble pas nécessaire de demander au comité d'évaluation une certification scientifique. Peut-être le comité peut-il en revanche, de son propre chef, donner une publicité particulière à certains des travaux qu'il juge d'une grande qualité ou particulièrement pertinents dans l'observation du comportement des plateformes.

De manière parallèle, il serait pertinent et souhaitable que les acteurs de la *peer review* puissent disposer du même accès aux données que les chercheurs sans quoi il n'est pas possible d'effectuer celle-ci convenablement.

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** ?

Concernant la question des données personnelles éventuellement utilisées pour produire des résultats de recherche, il convient évidemment d'appliquer les règles énoncées par le RGPD.

Concernant l'indépendance des chercheurs, il est important que les plateformes, une fois les données rendues accessibles via un moyen ou un autre, ne puissent pas décider *a posteriori* de contrôler les travaux de recherche effectués en les utilisant. Dans le cadre du projet *Social Science One,* par exemple, Meta s'octroie un droit de regard pré-publication afin de contrôler que les données utilisées ou les calculs présentés ne comportent pas de risques en matière de confidentialité de leurs utilisateurs. Cela ne devrait pas être leur rôle. Comme l'a montré l'affaire Timnit Gebru à Google, il arrive même que les propres chercheurs des plateformes voient la publication de leurs travaux refusée.

D. Protection des données et considérations techniques

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.) :

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

Si l'on veut obtenir plus d'accès à des données de la part des plateformes, il n'est pas imaginable que chaque demande d'accès suppose la création de bases spécifiques ou uniques. Un rôle du comité d'évaluation serait justement de faire remonter auprès des plateformes l'ensemble des métadonnées potentiellement utiles ainsi que les différents niveaux de granularité génériques nécessaires pour répondre aux différents besoins sollicités. Libre ensuite à chaque équipe de recherche de reformater et retravailler les données pour ses propres besoins.

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 ?

Comme l'illustrent les nombreux travaux déjà existants sur la base des données d'ores et déjà accessibles, il ne fait aucun doute que la mise à disposition de nouvelles données plus précises ou plus riches ne pourrait qu'alimenter ces travaux et susciter l'apparition de nouveaux indicateurs ou mesures innovantes.

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[1] ?

Oui à une co-construction, mais pas à part égale. Voir supra les réponses aux questions C.4. et C.5.

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 ?

Afin d'assurer à la fois une adaptation progressive et une culture commune pérenne, on pourrait envisager un comité à la composition assez large (une trentaine de membres a minima), renouvelé partiellement régulièrement (par exemple des mandats de 4 ans avec renouvellement tous les 2 ans par moitié).

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 ?

Dès lors qu'un accès a été autorisé pour un projet, il est important que celui-ci soit autant que possible semi-automatisé via une API ou un lien de téléchargement direct, plutôt qu'une mise à disposition discrétionnaire par échange électronique. Cela permet notamment d'assurer la possibilité de mettre à jour ou reconduire une étude avec des paramètres légèrement différents.

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** ?

Si les données sont disponibles de manière automatisée au travers d'une API, l'utilisation de clés d'API avec des authentifications fortes (type OAuth/OAuth2) disposant de droits d'accès distincts à différentes données et granularités pourrait suffire.

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

Les conditions de stockage de données prévues par le RGPD en matière de données personnelles s'appliqueront quoi qu'il en soit pour ces types de données.

[1]

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 ?

Il semble essentiel que ces différentes autorités soient directement impliquées dans les discussions menant à la définition des données mettables à disposition ainsi que des critères de validation des données transmises aux projets retenus, éventuellement avec un rapport annuel de contrôle soumis à l'approbation de ces autorités.

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 ?

Non, l'indépendance du comité vis-à-vis des équipes de recherche semble de rigueur.

Protections des données et considérations techniques : remarques complémentaires

E. Faisabilité de l'accès et incitations

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 ?

Donner visibilité et lisibilité aux processus permettant l'accès à ces données. Ces sujets sont techniques et complexes et, pour favoriser la mobilisation de nouvelles équipes de recherche issus de disciplines différentes, il est nécessaire de donner plus de publicité aux nouvelles possibilités qui vont s'ouvrir - *a contrario*, le dispositif *Social Science One* est méconnu par beaucoup de chercheurs.

Les établissements de recherche ayant des équipes et des laboratoires travaillant sur ces sujets - notamment les laboratoire de *computational social science* - pourraient constituer des centres d'expertise afin de favoriser la prise en main de ces données par des équipes de recherche d'autres disciplines ou ayant moins de familiarité avec ces données.

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 ?

Le comité d'évaluation des projets doit inscrire des critères de diversité des équipes dans ses objectifs (comme le font les comités de l'ANR). Par ailleurs, ces travaux ne pouvant être conduits sans financement, il est important que l'ANR développe un axe de recherche spécifique sur ces questions en étant attentif à la diversité des champs disciplinaires considérés (sciences sociales computationnelles, digital humanities, informatique, économie, sociologie, droit, etc.).

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 ?

Afin de bénéficier d'un effet de levier suffisant, il est d'abord nécessaire de favoriser une meilleure coordination des équipes de recherches et des régulateurs au niveau européens pour éviter les stratégies de demandes de données isolées et simplement nationales.

Pour inciter les plateformes, il nous semble qu'un dialogue de confiance s'appuyant sur la mise à disposition, la communication et l'échange autour des résultats des recherches peut créer une atmosphère favorable. Tous les résultats d'enquêtes ne sont pas critiques ou défavorables aux plateformes. Dans de nombreux cas, ce sont des recherches indépendantes qui ont montré que les moteurs de recherches ne produisaient pas la "bulle de personnalisation" souvent évoquée dans les médias. Ce sont des travaux de recherche qui montrent aujourd'hui que l'amplification algorithmique de contenus extrémistes sur YouTube (appelé rabbit hole) ne sont plus observables depuis le changement de la politique algorithmique et de modération de YouTube en 2019.

Un dialogue sérieux et constructif autour des résultats de recherche peut s'établir avec les plateformes. Il ne fait pas de doute qu'une meilleure ouverture des données aidera à diminuer le climat de méfiance qui s'est parfois installé entre les plateformes et la communauté académique.

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

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 ?

Pour ne pas compliquer les choses, cette tâche pourrait être confiée au Comité d'évaluation (voir *supra*).

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

Pour ne pas compliquer les choses, cette tâche pourrait être confiée au Comité d'évaluation (voir *supra*).

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

Nous n'avons pas de propositions en réponse à cette question.