

Arcom

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

CARISM



Consultation publique sur l'accès aux données des plateformes en ligne pour la recherche

Mai 2022

Consultation publique sur l'accès aux données des plateformes en ligne pour la recherche

1. L'accès aux données des plateformes pour la recherche : un enjeu central dans un monde en transformation

1.1. Les évolutions récentes des réseaux sociaux et des usages en ligne redéfinissent nos modes d'accès à l'information

Les moteurs de recherche, les plateformes de partage de vidéos et les réseaux sociaux redéfinissent la façon dont les contenus, notamment d'information, sont consommés et partagés.

Ces sources d'innovation ont débouché sur de **nouvelles voies d'expression et ont accéléré certaines formes de participation citoyenne**. Toutefois, elles peuvent également être l'objet de détournements et dérives. Parmi celles-ci, on compte notamment les phénomènes de manipulation de l'information ou de haine en ligne.

L'environnement informationnel actuel ne se définit ainsi plus par l'addition de secteurs dont les frontières seraient hermétiques : audiovisuel et numérique ; médias traditionnels (télévision, radio, presse) et nouveaux services de consommation de contenus (réseaux sociaux, applications) ; modes de réception historiques et terminaux de demain ; médias nationaux, européens et internationaux. Les recoupements sont au contraire désormais de plus en plus importants. Ils donnent lieu à des phénomènes de redistribution des temps d'attention consacrés aux médias et des sources choisies, qui renforcent le **rôle structurant et croissant d'internet dans l'accès à l'information**. **Les usages sur internet rivalisent à présent avec ceux des médias traditionnels**¹.

À ce rôle d'accès à l'information s'ajoute également un effet d'internet en général, et des réseaux sociaux en particulier, sur la formation des opinions. Une exposition renforcée à des contenus proches ou similaires aux opinions connues des utilisateurs constitue par exemple l'une des caractéristiques principales des fils d'actualité sur les réseaux sociaux.

1.2. Le monde de la recherche a un rôle déterminant à jouer dans la compréhension des usages en ligne

Dans ce contexte, **il est crucial que la recherche soit en mesure d'étudier ces nouvelles dynamiques et de développer des outils et approches indépendants afin de les éclairer**. Il en va en effet de la connaissance collective de phénomènes dont les effets potentiels peuvent être délétères sur nos sociétés.

L'élaboration d'un cadre permettant l'étude des comportements en ligne et leurs effets, doit contribuer à la **protection et au renforcement de l'indépendance, de l'autonomie et de la capacité d'analyse** propres à la recherche, et lui permettre de jouer son rôle dans l'accompagnement et la compréhension des changements sociétaux contemporains.

¹ Selon le dernier baromètre médias Kantar/La Croix, les Français et Françaises placent internet comme deuxième moyen d'information (32 %) derrière la télévision (48 %) mais devant la radio (13 %) et la presse écrite (6 %). Néanmoins, la confiance accordée à ces différents supports n'est pas corrélée positivement à leurs usages : ainsi, la radio et la presse écrite sont considérées comme les moyens d'informations les plus fiables à 49 %, juste devant la télévision (48 %). De ce point de vue, les médias traditionnels conservent encore et largement la confiance de de leurs usagers. À l'opposé, seuls 24 % des Français estiment qu'on peut trouver des informations crédibles sur internet.

Il convient donc de mener une réflexion sur le rôle que peut jouer la puissance publique pour aider le monde de la recherche à se saisir pleinement de ces problématiques. Ce rôle de facilitateur doit plus particulièrement s'exprimer dans **l'exploitation et l'analyse des données issues des réseaux sociaux ou des services de plateformes en ligne et qui conditionnent le développement des connaissances propres aux environnements numériques**. L'enjeu de la bonne exploitation de ces données est double : il s'agit à la fois de pérenniser un écosystème de recherche dynamique, effectif et durable, capable de générer de la connaissance au bénéfice de tous (**production scientifique**), mais également de contribuer à l'expertise du régulateur dans son évaluation des dispositifs mis en œuvre par les opérateurs de plateformes pour satisfaire à leurs obligations telles que de la modération des contenus haineux (**régulation de la transparence**).

2. Pourquoi l'Arcom entend jouer un rôle dans l'accès aux données des plateformes pour la recherche

2.1. Dans le respect du RGPD, le régulateur doit être un facilitateur dans l'accès aux données pour le monde de la recherche

Née de la fusion du Conseil supérieur de l'audiovisuel (CSA) et de la Haute Autorité pour la diffusion des œuvres et la protection des droits sur internet (Hadopi) le 1^{er} janvier 2022, **l'Autorité de régulation de la communication audiovisuelle et numérique (Arcom) a été créée afin d'accompagner les importantes transformations du paysage audiovisuel et numérique**. La régulation est une des réponses apportées à ces défis bien identifiés par la puissance publique. L'Arcom est notamment chargée de protéger la création et ses acteurs, veiller aux équilibres économiques du secteur audiovisuel, superviser les moyens mis en œuvre par les plateformes en ligne pour protéger les publics tout en garantissant la liberté d'expression, et assurer le pluralisme politique sur les antennes. Son action vise plus largement à protéger tous les publics dans l'audiovisuel et en ligne.

De plus, les pouvoirs de régulation systémique des opérateurs de plateformes en ligne (comme définis par l'article L. 111-7 du Code de la consommation) **confiés à l'Arcom par le législateur se sont renforcés de manière continue depuis fin 2018**. Ils s'appliquent principalement aux réseaux sociaux (Facebook, Snapchat, etc.), aux moteurs de recherche (Google, Bing, etc.) et aux plateformes de partage de vidéos (Dailymotion, YouTube, etc.). C'est dans une acception large des « plateformes » que l'Arcom entend ici mener cette consultation, afin d'ouvrir le débat à l'ensemble des acteurs de l'écosystème informationnel numérique, pour englober de nouvelles catégories d'acteurs qui pourraient émerger dans le court ou le moyen terme et rentrer dans la catégorie des « plateformes ».

Ce nouveau paradigme, qui vient compléter son modèle de régulation, donne à l'Arcom une nouvelle place au sein d'un écosystème étendu et polymorphe. L'Autorité supervise les moyens mis en œuvre par les opérateurs, lesquels ont un devoir de coopération et de transparence². Le monde de la recherche conduit des travaux afin d'éclairer la compréhension de ces phénomènes. La société civile dans son ensemble participe à ces actions par ses analyses, ses retours d'expériences et ses alertes. Ces différents champs d'action se complètent et forment une **boucle de rétroaction** où le régulateur est un acteur aux côtés d'autres pour identifier, analyser, évaluer, questionner puis au besoin, proposer des

² Dans les limites qui doivent être dûment justifiées par exemple au titre de la sécurité de leurs services.

mécanismes de réponse aux risques identifiés. **Il est également important de souligner que cette démarche s'inscrit dans le cadre juridique européen du règlement général sur la protection des données personnelles (RGPD) des utilisateurs des services de plateformes en ligne.** Le RGPD a vocation à s'appliquer à une très grande majorité des traitements de données personnelles mis en place par chacun des acteurs. L'anonymisation des données issues des plateformes étant techniquement complexe à mettre en œuvre en pratique et pouvant avoir des effets sur la définition des questions de recherche, la bonne prise en compte de ce caractère personnel des données est d'importance. La CNIL a d'ailleurs conduit une consultation publique auprès des chercheurs quant à leurs modes d'accès aux données et au regard du RGPD. Cette initiative a débouché sur la **publication de ressources pour ces acteurs** : présentation des enjeux et règles à respecter, rappel des outils à disposition pour la mise en conformité, etc.³. **Les problématiques d'accès aux données sur les plateformes en ligne s'inscrivent donc dans ce cadre de protection des droits des utilisateurs, notamment du droit à la maîtrise** des données par les personnes concernées⁴.

2.2. Les pratiques actuelles des opérateurs de plateformes en ligne en matière d'ouverture de leurs données sont très diverses

Pour que l'ensemble des acteurs puissent jouer leur rôle, il est nécessaire que l'identification des problématiques qui se posent sur les services de plateformes en ligne ne repose pas sur les seules initiatives des opérateurs. Au-delà de ce que ces acteurs rendent disponible, au demeurant de manière très hétérogène, le monde de la recherche doit pouvoir également accéder à des données de qualité selon des modalités qui ne soient pas définies par les plateformes seules. C'est ainsi **une régulation de la transparence qui doit être déployée**, dans laquelle l'Arcom doit pouvoir se nourrir des apports des différents acteurs tout en ayant un rôle de mise en capacité de ces parties prenantes à agir.

En effet, l'accès aux données des plateformes en ligne est aujourd'hui complexe, notamment du fait de l'absence de cadre unifié ou de politique de mise à disposition commune entre les plateformes, au niveau national ou supranational. Cet état de fait est notamment souligné par des initiatives telles que l'*European Digital Media Observatory* (EDMO)⁵. Créé en 2020 et mené principalement sous l'égide de l'*Institut Universitaire Européen* de Florence (EUI), ce groupe d'experts venus du milieu universitaire, du secteur des médias ou d'instances gouvernementales vise à apporter de nouveaux éclairages sur les questions de désinformation en ligne. Dans cette perspective, l'EDMO a au titre de ses objectifs de contribuer à la réflexion sur l'utilisation des données des plateformes en ligne **notamment en soutenant les autorités compétentes dans leurs démarches de régulation**⁶.

³ <https://www.cnil.fr/fr/recherche-scientifique-hors-sante>

⁴ L'identification des rôles et des responsabilités de chaque acteur au regard du RGPD, notamment au regard de la transparence due aux personnes concernées doit permettre aux personnes d'exercer leurs droits. Cf. « [air2021 : entre partage et protection, quelle éthique pour l'ouverture des données ?](#) », CNIL

⁵ <https://edmo.eu/>

⁶ Le deuxième objectif qui apparaît dans le rapport d'activité de l'EDMO de 2021 est le suivant : « Creating a governance body which ensures public trust regarding the work of the platform and establishing a framework to provide secure access to data of online platforms for research purposes ». (Source : <https://edmo.eu/wp-content/uploads/2021/09/EDMO-Public-Report-June-2020-%E2%80%93-March-2021-2021.pdf>)

Les accès sont aujourd’hui majoritairement permis par les plateformes de manière volontaire, concentrant les recherches sur les services les plus allants en la matière.

S’il faut saluer ces initiatives, force est de constater que les recherches se concentrent surtout sur Twitter, qui offre différentes API dont une dédiée à la recherche⁷. Cette ouverture a permis à de nombreux projets de voir le jour, notamment via la collecte automatisée de contenus. À titre d’illustration, l’on peut citer l’initiative de l’*Institut des Systèmes Complexes de Paris Ile-de-France* (ISC-PIF, laboratoire CNRS) qui réunit depuis 2016 une équipe de chercheurs et d’ingénieurs pour exploiter les données de ce réseau social. Le travail scientifique de traitement et d’analyse des données a par exemple permis la mise en œuvre du *Politoscope*⁸ : cet outil de visualisation à destination du grand public a pour but d’éclairer les thèmes qui font l’actualité politique et leurs évolutions⁹. **D’autres réseaux sociaux ou moteurs de recherche font le choix d’une politique d’accès à leurs données plus restrictive, y compris pour les chercheurs.**

2.3. L’Arcom se positionne au cœur des réflexions ouvertes par le *Digital Services Act* (DSA), qui traite des enjeux les plus actuels tout en soulevant des questions opérationnelles

Pour répondre aux enjeux portés par les plateformes en ligne, la nécessité d’une action au niveau européen s’est progressivement imposée. Celle-ci s’exprime notamment par la prise en considération des problématiques relatives à l’émergence et à la consolidation de nouveaux marchés numériques, avec le *Digital Markets Act* (DMA), et de celles autour de la circulation des données entre entreprises, avec le *Data Governance Act*.

À ces initiatives s’ajoute le *Digital Services Act* (DSA) ; cette proposition de législation européenne vise à garantir la sécurité des utilisateurs et la protection des droits fondamentaux en ligne. L’Arcom, à travers notamment plusieurs prises de position de l’ERGA, accueille très favorablement cette évolution de la régulation. Le DSA propose notamment un modèle de **régulation systémique** des plateformes en ligne de nature à répondre à certains des désordres informationnels les plus importants de notre époque tout en préservant l’une des caractéristiques intrinsèques d’internet, offrir un espace d’exposition et d’expression. Pour les très grandes plateformes en ligne¹⁰, des obligations supplémentaires sont prévues afin d’augmenter encore le niveau de transparence de leur action, notamment en matière de fonctionnement de leur modération, de leurs services publicitaires et des algorithmes qu’elles utilisent sur leurs services.

Plus spécifiquement, **l’article 31 du DSA** vise à encadrer l’accès des chercheurs aux données de ces très grandes plateformes afin de contribuer à l’évaluation des risques systémiques que leurs services peuvent présenter. Le DSA se place dans une perspective de renouvellement de

⁷ Il faut cependant noter que plus généralement en termes de recherche, les plateformes peuvent conduire en interne des travaux ou mandater directement des chercheurs externes. Ces initiatives restent à la discrétion des acteurs et ne supposent pas la création de dispositifs pérennes d’accès à des données.

⁸ *Projet Politoscope, CNRS Institut des Systèmes Complexes Paris Ile-de-France* (ISC-PIF), <http://politoscope.org>

⁹ L’exemple du *Politoscope* n’a aucunement vocation ici à servir de modèle de dispositif de recherche qui aurait la préférence de l’Arcom : il est ici utilisé afin d’illustrer comment la collecte automatisée de données d’un réseau social a donné lieu à une exploitation scientifique qui a généré une contribution au débat public sous la forme d’un outil mis à disposition du public.

¹⁰ La catégorie des « très grandes plateformes en ligne » (*very large online platforms* ou *VLOP*) englobe les services qui touchent au moins 45 millions d’utilisateurs dans l’Union européenne par mois. Voir notamment : « Digital Services Act Briefing », *European Parliament*, 2021. URL : [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689357/EPRS_BRI\(2021\)689357_EN.pdf#:~:text=The%20DSA%20proposal%20is%20a%20horizontal%20instrument%20putting,and%20Digital%20services%20act%20%28DSA%29%20draft%20asymmetric%20obligations](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/689357/EPRS_BRI(2021)689357_EN.pdf#:~:text=The%20DSA%20proposal%20is%20a%20horizontal%20instrument%20putting,and%20Digital%20services%20act%20%28DSA%29%20draft%20asymmetric%20obligations)

la relation entre les plateformes, les autorités et les usagers et pourrait aboutir à l'émergence d'un **nouveau modèle de régulation**¹¹. Ainsi, le monde de la recherche serait étroitement associé à la meilleure compréhension des **dynamiques socio-économiques, politiques et culturelles** qui émergent dans ce nouvel écosystème informationnel. L'Arcom espère contribuer à son échelle à la réflexion sur ces questions d'accès et de construction d'un modèle innovant au niveau européen.

L'article 31 du DSA soulève toutefois la question de sa pleine opérationnalité au vu des objectifs poursuivis :

- La place de l'intermédiaire entre chercheurs et plateformes : le « *coordinateur de l'État membre d'établissement* » (*Digital Services Coordinator*) est l'un des deux intermédiaires, avec la Commission, entre les parties prenantes. La définition de son rôle sera donc particulièrement structurante.
- Les données concernées par cet accès : le champ des données visé est celui de « *l'identification et [de] la compréhension des risques systémiques* » au sens du DSA. Ces risques devraient, dans l'état actuel des textes, recouvrir trois catégories en particulier : les potentielles manipulations des services de plateformes notamment pour diffuser des contenus illégaux ou pour des finalités économiques ; l'impact de ces services sur les droits fondamentaux comme la liberté d'expression eu égard notamment aux systèmes algorithmiques utilisés ; et les manipulations intentionnelles afin de diffuser massivement des informations pouvant avoir un impact délétère sur la santé publique, les processus électoraux ou la sécurité. Il faut se féliciter que ces champs couvrent les problématiques les plus urgentes parmi les désordres informationnels déjà identifiés par la recherche. Néanmoins, peut se poser la question de la pertinence d'une approche plus englobante, en particulier dans une perspective de recherche interdisciplinaire. De plus, il reste capital d'être en mesure d'identifier de nouveaux risques dans le futur et qui ne seraient pas encore observés à l'heure actuelle mais que la recherche pourrait identifier.
- Le statut des chercheurs autorisés à accéder à ces données : l'article 31 conditionne cet accès à certains critères. Cette disposition donnerait ainsi un cadre clair aux chercheurs qui souhaiteraient étudier les phénomènes couverts par le DSA, sans préjudice du RGPD. Les futurs actes délégués pourront préciser les conditions dans lesquelles de tels accès seraient fournis aux chercheurs qui en feraient la demande. Il semble utile ici de s'interroger quant au risque que des critères trop stricts (capacités administratives ou financières de la structure demandeuse, travaux relatifs précédemment menés par un ou des membres de l'équipe de recherche, possibilités effectives d'interdisciplinarité, etc.) dans les conditions d'éligibilité à des accès des données ou dans les projets retenus pourraient avoir des effets de bord limitants. Par exemple, la valorisation de l'expertise technique pourrait favoriser largement les chercheurs universitaires ayant déjà produit de nombreux articles sur les sujets visés par le DSA. C'est donc tout un continuum de recherche qui pourrait être mis à l'écart des dispositifs d'accès : jeunes chercheurs, journalistes, ONG, etc. Cette question soulève également celle de l'éventuel arbitrage entre ouverture à un large nombre d'acteurs et les risques en termes d'utilisation de données personnelles pour les

¹¹ Sur les ambitions du DSA et ses possibles répercussions sur le débat international autour de la régulation des plateformes et de l'organisation de la transparence, voir par exemple Schiffrin (2021), qui souligne les résonnances que pourrait avoir le DSA aux États-Unis : https://www.cjr.org/business_of_news/europe-regulates-big-tech.php

personnes concernées. La qualification de la recherche scientifique au sens du RGPD peut en effet s'avérer plus restrictive qu'une évaluation strictement scientifique des projets.

2.4. L'Arcom entend se placer dans un cadre ouvert et contributif pour établir le modèle d'accès aux données des plateformes en ligne.

C'est dans ce cadre que l'Arcom lance la présente consultation publique sur l'accès aux données des plateformes en ligne pour la recherche et en lien avec les problématiques sur lesquelles l'Autorité a compétence : lutte contre la manipulation de l'information et haine en ligne.

A travers cinq thèmes – partage d'expériences d'utilisations de données de ces services (A), gouvernance (B), construction des projets scientifiques (C), protection des données et considérations techniques (D), et faisabilité des accès et incitations (E) – cette consultation publique vise à interroger l'ensemble des parties prenantes. Il s'agit de tirer de premiers enseignements quant à la mise en œuvre d'un cadre opérationnel d'accès aux données de plateformes en ligne et de contribuer ainsi à la réflexion générale des différentes parties prenantes sur ces problématiques, en particulier les chercheurs et la sphère publique. Monde académique, plateformes en ligne, pouvoirs publics et associations sont ainsi invités à partager leurs idées et contribuer à l'intérêt général au travers de la recherche.

Les éléments recueillis par l'Arcom feront ensuite l'objet d'une synthèse qui visera à nourrir les débats déjà existants en matière d'accès de la recherche aux données des plateformes en ligne ; ce travail pourra susciter le cas échéant de nouvelles réflexions aux niveaux français, européen et international. L'ensemble des réponses ainsi que la synthèse seront rendues publics¹².

Les contributions à la consultation doivent parvenir à l'Arcom avant le 22 juillet 2022 à l'adresse électronique suivante : consultation@arcom.fr

3. L'Arcom entend nourrir sa réflexion sur la base des réponses à cinq grandes thématiques de questions

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

- *Questions à destination de tous les acteurs intéressés par l'étude et la recherche en lien avec les plateformes en ligne :*

L'intérêt pour les questions relatives aux plateformes et l'étude des activités en ligne ont intégré l'agenda de recherche d'un nombre croissant de disciplines. Ces champs d'études sont variés, allant des **sciences de la nature à l'informatique en passant par les sciences sociales**. Ils impliquent de ce fait un traitement de la donnée s'appuyant sur des **protocoles et méthodologies** variés et nécessitent de prendre en compte les éventuelles spécificités disciplinaires qui rendraient certaines modalités d'accès et d'étude plus appropriées que d'autres selon les questions de recherche. De plus, certains services ont **une politique**

¹² La publication des réponses à des fins de transparence n'exclut toutefois pas la possibilité pour les répondants de demander à ce que certaines de leurs réponses soient traitées de manière confidentielle.

d'ouverture de leurs données aux chercheurs, notamment via la mise à disposition d'API, tandis qu'à l'inverse l'accès peut être restreint voire soumis à un contrôle strict chez d'autres.

Les questions suivantes visent à mieux appréhender les **expériences qu'ont pu avoir les répondants dans leurs projets de recherche avec les données des plateformes**, les **difficultés** auxquelles ils ou elles ont pu faire face, et les éventuelles **contraintes** d'ordre technique ou légal qui auraient influencé la construction de leurs recherches.

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.) ?

Les chercheurs du CARISM ayant répondu au questionnaire ont conduit plusieurs types de recherches mobilisant des données des plateformes :

i) des **analyses de débat en ligne** au moyen de commentaires et de fils de discussion en ligne.

Pour ces recherches, la collecte de données est différente selon les plateformes :

- Les données Facebook sont récoltées « à la main », copiées-collées dans Excel, et/ou prises en capture d'écran à l'aide d'outils comme GoFullPage. Des logiciels de collecte existent mais ils sont a) très chers et donc hors de portée de la plupart des chercheurs et de leurs laboratoires ; et b) limités aux groupes Facebook publics. Il peut aussi être compliqué d'avoir recours à de tels outils sans savoir si leur utilisation est totalement légale.
- Les commentaires YouTube peuvent être récoltés à l'aide du logiciel YouTube Data Tools
- Les tweets à partir de l'API Twitter et à travers le logiciel Minet du Médialab. L'outil TwitterArchiver est aussi utile

ii) des recherches sur les **contenus « problématiques » (terroristes, haineux, de désinformation) et leur modération** sur Facebook, YouTube et Twitter. Pour ces travaux, on s'appuie principalement sur des données publiées par les plateformes elles-mêmes : leurs rapports de transparence, conditions d'utilisation, communications officielles, etc. Ces données étant partielles et partiales, elles nécessitent généralement d'être complétées par des entretiens avec des acteurs (représentant les plateformes, faisant de la veille sur leurs services ou visant à les réguler) afin de comprendre les catégories et les métriques présentées.

iii) des recherches sur les **algorithmes de recommandation**, notamment celui de Netflix. Ici, la collecte a été faite manuellement par le chercheur qui a noté à la main dans un fichier excel les mots clés liés à chaque film, les tags associés, et de manière quotidienne, leur place dans les différentes catégories proposées sur l'interface. Pour comparer différents profils utilisateurs, plusieurs comptes ont été créés et les étapes précédemment décrites répétées en utilisant un VPN et différents ordinateurs.

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

La méthode de collecte manuelle de contenus et de commentaires, notamment sur Facebook, est chronophage et fastidieuse. Les chercheurs concernés mettent en avant la difficulté de remonter dans le temps sur les groupes Facebook pour collecter les commentaires jusqu'à la date qu'ils s'étaient fixée (la fonction « rechercher » dans les groupes est limitée puisqu'elle ne permet pas de recherche par date). L'un d'eux a pu, par la suite, utiliser des logiciels qui permettent d'aspirer les données de groupes Facebook publics et d'extraire les discussions sous forme de fichier Excel. Ces outils présentent un véritable gain de temps mais sont très chers, donc pas à la portée de tous les chercheurs. Il serait utile d'avoir une API permettant d'exporter des discussions depuis des groupes Facebook publics, en fonction de dates ou de mots clés spécifiques.

Pour ceux qui travaillent sur les questions de modération de contenus : les rapports de transparence publiés par les plateformes sont incomplets, d'une régularité fluctuante et parfois carrément trompeurs. Pour donner un exemple très précis, cela fait des années que Facebook affiche un taux de détection automatique des contenus terroristes de 99% (c'est-à-dire que 99% des contenus supprimés pour ce motif ont été détectés par ses outils de modération automatisés), sans stipuler que son algorithme ne fonctionnait que sur deux groupes (Daech et al-Qaïda). En clair, il faudrait pouvoir avoir accès à des rapports harmonisés (chaque plateforme remplit le même pour que l'on puisse comparer) et dont les métriques, voire même les catégories, ne sont pas définies par les plateformes elles-mêmes. On note aussi que certaines pratiques de modération de plus en plus mobilisées par les plateformes ('soft moderation') échappent aux rapports de transparence qui ne parlent que de contenus signalés/supprimés. Il faudrait aussi que les plateformes soient tenues de fournir des statistiques sur les pratiques de *shadowbanning* (à quel point la visibilité d'un post est dégradée, réduction du nombre de vues, nombre de contenus concernés, etc.)

De manière plus générale, on peut préciser que la majorité des données sur lesquelles travaillent les chercheurs en information communication sont « disponibles », sous-entendu visibles (profil d'utilisateur, heures et dates, contenus publiés, échanges dans un groupe, commentaires etc.), mais disséminées, cachées, et parfois supprimées (shadowban, groupes privées, résultats propres à un profil particulier, etc.). Il faut utiliser des moyens détournés pour y accéder, ce qui demande un investissement très important en temps et en ressources machines. De plus, les plateformes luttent parfois activement contre l'utilisation de tels moyens détournés (par exemple, créer un faux compte, utiliser un VPN), ce qui ajoute une part d'incertitude quant à notre capacité à aller au bout de certains projets de recherche.

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.

Sur Facebook, certaines de nos recherches ont exclu par défaut les groupes privés du fait de l'impossibilité d'y accéder (pas d'admission par les administrateurs) ou de collecter les données issues de ces groupes (les logiciels de collecte n'y ont pas accès).

Sur Twitter, l'API ne permettant d'avoir accès qu'à un certain pourcentage des tweets sur un hashtag en particulier, certaines recherches sur des événements très commentés et qui

visent l'exhaustivité sont abandonnées. De même, les recherches « historiques » sont impossibles via l'API (seuls les tweets de moins d'une semaine sont pris en compte).

L'interdiction de plus en plus concrète des « faux comptes » a aussi conduit à la réorientation d'un projet de thèse sur les menaces à la sécurité nationale sur les réseaux sociaux. Il est en effet nécessaire pour les chercheurs qui souhaitent étudier les mécanismes de suggestion algorithmique, ou faire de la veille sur des sujets sensibles de ne pas utiliser leurs comptes personnels. Or à cause de l'efficacité des mécanismes de détection automatique anti-spam ou anti-faux comptes, entretenir un faux compte est devenu très compliqué : il faut adopter un comportement d'utilisateur « normal » : ajouter des amis, poster des contenus, en « liker » d'autres, ajouter une photo de profil d'un humain, etc. tout ça sans certitude que le compte en question (et donc tout le travail prévu) ne sera pas supprimé ou suspendu pour une période indéterminée si un faux pas est commis. Par exemple, une doctorante en info-comm de l'université de Cergy-Pontoise a dédié sa thèse au suivi de communautés de Daech sur Facebook. Pour des raisons évidentes, elle n'utilise pas son compte personnel pour son travail, mais son faux compte a été supprimé à plus de 100 reprises par Facebook. Il serait donc extrêmement utile de pouvoir bénéficier de "faux comptes" ou comptes fantômes de chercheur, que l'on pourrait par exemple faire enregistrer auprès des plateformes pour qu'elles ne les comptabilisent pas (pour ne pas nuire à l'intégrité de leurs plateformes).

De manière plus générale, il faut souligner que les chercheurs qui travaillent sur les plateformes ont conscience des possibilités, ou plutôt, de l'absence de possibilités offertes, celles-ci sont prises en compte dans la conception même des projets de recherche. De nombreux projets ne sont même pas envisagés car impossibles à mener à bien, soit parce que les données ne sont pas accessibles, soit parce que leur traitement demande un investissement trop important. Une forme d'autocensure existe : pourquoi lancer des projets impossibles à réaliser ? Par exemple, c'est sans doute la disponibilité d'une API chez Twitter qui explique pourquoi cette plateforme est la plus étudiée, alors même que l'on sait que les utilisateurs de Twitter sont peu représentatifs de la population, et que cette plateforme a nettement moins d'utilisateurs que ses concurrents chez Meta, YouTube, TikTok, etc.

De plus, nous n'avons pas connaissance de tous les aspects mesurés par les plateformes à propos des conduites individuelles et des contenus qu'ils gèrent.

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 aucun cas nous avons bénéficié de l'appui des plateformes étudiées, mais on peut souligner l'avance de Twitter en la matière, via son dispositif *Twitter for developers* qui permet, après avoir expliqué brièvement son projet de recherche, d'accéder à une API. Un dispositif similaire pour Facebook, YouTube, Snapchat, Google, Instagram, Dailymotion ... serait très utile.

➤ *Questions spécifiques à destination des plateformes en ligne :*

Les politiques de mise à disposition des données à destination de la recherche diffèrent sensiblement d'une plateforme à l'autre. Les questions suivantes visent à mieux appréhender **leurs politiques respectives** et à en comprendre les déterminants : nature du service, spécificités techniques ou juridiques, ou encore évaluation de risques spécifiques que le partage de données pourrait poser.

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.

A. Partage d'expériences d'utilisations de données des services en relation avec la thématique : remarques complémentaires

B. Gouvernance

➤ *Définition des acteurs :*

L'accès à des données utiles à la société pose la question de leur **ouverture à tous les acteurs** de la recherche. Si le monde académique semble être le principal bénéficiaire d'un accès plus ouvert, la contribution **des think tanks, des journalistes et de la société civile** à la connaissance des problématiques liées aux plateformes en ligne mérite réflexion¹³. La question de **la neutralité des acteurs**, au vu des financements qu'ils peuvent notamment recevoir de certaines plateformes, se pose également.

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) ?

Oui, des limites paraissent nécessaires pour s'assurer que les projets ne suivent pas des intérêts commerciaux mais uniquement des visées scientifiques et publiques. Les critères retenus pour obtenir l'accès aux données pourraient être de plusieurs ordres selon la nature de l'acteur et de son projet.

Pour la recherche, le projet devrait être validé par une école doctorale, un laboratoire ou un département universitaire, jugeant de sa pertinence (intérêt scientifique, particularités, originalité, sérieux, etc.). Le système d'accréditation pourrait ressembler à ce qui est déjà en place pour la BNF ou l'INA par exemple. Dans la proposition soumise, le type de données nécessaires pour mener à bien l'étude devrait être clairement explicité afin de pouvoir configurer des points d'accès spécifiques.

Pour un projet journalistique, la démarche serait identique via des associations ou des institutions existantes telle l'Arcom. Également, un groupe pourrait être créé sur le modèle du Conseil de déontologie journalistique et de médiation, il ferait le lien entre la grande diversité de médias (*pure players*, traditionnels, tv, radio, presse, grands public, spécialisés), leurs projets et les plateformes. Les journalistes amateurs/citoyens, sans carte de presse, qui mènent par exemples des enquêtes type OSINT, devraient aussi pouvoir soumettre leurs propositions (voir groupe Bellingcat).

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

Les possibilités d'accès doivent être différentes, bien que tous doivent pouvoir accéder aux mêmes données. Cette différence s'explique notamment par le temps nécessaire à la réalisation de chacun des projets. La recherche s'inscrit dans un temps long d'observation et de comparaison. De plus, les sujets de sciences sociales et d'information communication peuvent être multiples et extrêmement précis, allant de

¹³ L'une des modalités de ces contributions est par exemple la science et la recherche participatives. Il s'agit de « formes de production de connaissances scientifiques auxquelles participent, aux côtés des chercheurs, des acteurs de la société civile, à titre individuel ou collectif, de façon active et délibérée » à toutes les étapes du continuum de la recherche, comme par exemple la collecte de données, leur analyse et l'interprétation des résultats (Source : [La recherche participative · Inserm, La science pour la santé](#)).

la réception des série TV US chez les jeunes, à l'évolution des discours anti-avortement dans la région Rhône Alpes Auvergne. Il faut donc que les données puissent être anonymisées, conservées dans le temps, sans être complètement dénaturées de leur contexte d'existence.

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, ce point devrait faire partie de leurs obligations de transparence, à la manière des rapports qu'ils remettent à l'ARCOM, avec davantage de données chiffrées. Aussi, une interface générique pourrait être créée afin de permettre au grand public de se représenter les données possédées par les plateformes, sur quelles conduites elles portent, ainsi que leur quantité.

➤ *Modalités d'attribution d'accès aux données :*

Les modalités d'attribution des accès et les éventuels **critères** sur lesquels les projets de recherche seraient sélectionnés sont également à prendre en compte. En effet, si la légitimité de l'utilisation de données à des fins de recherche n'est pas en débat ici, la mise en application de ce principe soulève de nombreux enjeux. Les **rôles respectifs des institutions européennes ou nationales** qui pourraient être impliquées dans la sélection de projets de recherche est par exemple à considérer.

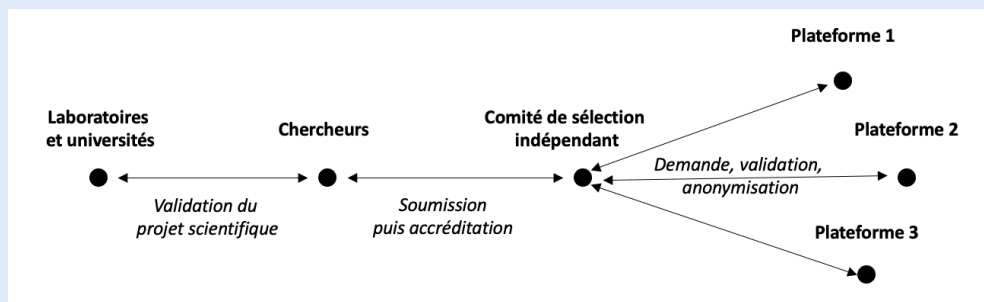
B.3. Selon vous, un modèle d'accès à des données via la formulation des demandes d'accès à un **tiers de confiance** est-il pertinent ?

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 ?

Le tiers de confiance se doit de connaître les modes de fonctionnement des différents acteurs qui souhaiteraient obtenir des données. Les particularismes et problématiques étant nationales, il ne paraît pas opportun de créer une supra-entité chargée de l'accès au niveau européen. De plus, elle se trouverait rapidement confrontée aux représentants nationaux des plateformes.

Le protocole, pour la recherche, pourrait être le suivant :



Le chercheur devrait soumettre son dossier à l'université ou à un laboratoire afin que son intérêt scientifique soit validé. Une fois l'accord obtenu, il est adressé au comité de sélection chargé de trier ce qui est ou non faisable avec les plateformes. Une fois cette étape validée, il accrédite le chercheur. C'est cette entité qui gère les échanges avec les plateformes, en s'assurant qu'elles répondent dans les temps et correctement au projet (qualification, catégories, métriques harmonisées, etc.). Dans cette configuration, les plateformes n'ont pas connaissance de l'entité qui requiert les données. Les informations transmises aux chercheurs sont anonymisées. Nous pouvons imaginer qu'elles sont aussi datées, c'est-à-dire qu'elles respectent un décalage temporel de par exemple six mois entre le moment où elles sont créées et où elles sont transmises. On comprend par là que la responsabilité du comité est autant envers les chercheurs qu'envers les plateformes.

- 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 rajout d'un tiers de confiance peut sensiblement allonger le processus et rajouter une lourdeur dans le traitement des projets. Dans le cas où serait privilégié des formes d'interactions directes entre plateforme et chercheurs, une piste de réflexion pourrait consister en la création de « chambres froides », lieux spécifiques dont les plateformes auraient le contrôle, et dans lesquels les chercheurs devraient se rendre pour accéder aux données. Ce modèle d'interaction existe pour l'accès aux archives audiovisuelles de l'INAthèque, et se compose des étapes suivantes : demande d'accréditation, déplacement physique aux archives, demande d'accès à certaines parties, réception via documentalistes.

Si l'option sans intermédiaire est privilégiée, les entreprises numériques devraient recruter des personnes qualifiées dédiées à l'accompagnement du travail des chercheurs, associations, journalistes, ... Aussi, un processus moins encombrant consisterait en la mise à disposition de profils au statut particulier (« chercheur », « journaliste », « association »), similaires à celui d'administrateur de groupe, sur les plateformes. Étant identifiés comme « non-utilisateurs », ou « utilisateurs fantômes », ils pourraient requérir sur demande des données et effectuer des études *in situ*, même sans possibilité d'interaction (limité aux seules publications publiques).

Enfin, certaines conditions devraient être imposées, notamment par les pouvoirs publics. Il est nécessaire par exemple qu'une harmonisation existe à propos du type de données

(dénomination, catégories, descripteurs, ...) disponibles pour chaque plateforme et de leurs métriques, afin de pouvoir effectuer des comparaisons.

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 ?
- 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 ?
- iii) comment la **transparence des décisions** des organisateurs du protocole d'accès devrait-elle être garantie ?
- iv) quelle place et quels rôles devraient avoir chacune des **parties prenantes** et notamment les plateformes ?
- v) identifiez-vous des **risques inhérents** à ce modèle ? Lesquels ?

C'est l'État qui doit s'assurer que le tiers de confiance respecte les engagements pour lesquels il a été créé. En l'occurrence l'Arcom semble toute désignée pour cette mission. Une commission pourrait se réunir deux fois dans l'année pour juger du fonctionnement, du nombre de projets traités, et de la bonne entente avec les plateformes quant aux données échangées.

Concernant la transparence des décisions, elle répond à un cahier des charges établi en amont et dont les critères sont disponibles publiquement ou sur demande. Les projets qui se voient refuser l'accès aux données doivent se situer en dehors du cadre préalablement établi. Par exemple, si un projet ne précise pas clairement l'intérêt qu'il a à demander tel type de données, pour quels modes de traitement, et dans quel but, le chercheur pourra se voir refuser sa demande. Tout refus doit être signifié sur des bases claires.

Les plateformes doivent pouvoir exprimer les limites de leurs compétences (humaines, technologiques), ou expliciter le temps nécessaire pour obtenir tel ou tel type de données. Globalement, elles ne doivent pas avoir de rôle dans la sélection des projets, sans quoi la recherche va se trouver orientée et biaisée.

Un point important réside dans la rapidité et la fluidité de la procédure d'accréditation et d'accès aux données, les chercheurs ayant un temps de plus en plus limité pour mener à bien leurs travaux (financements sur projet, contrats courts de type *postdoc*, contrats doctoraux de 3 ans, etc.). Pour reprendre le cas de certaines archives, de nombreuses recherches ont dû être abandonnées faute d'obtention d'une accréditation à temps. Le tiers de confiance ne doit pas alourdir le processus mais s'assurer que les plateformes font ce qui est nécessaire pour que les projets puissent exister.

B. Gouvernance : remarques complémentaires

C. Construction des projets scientifiques

Les transformations récentes et à venir des plateformes en ligne soulèvent la question de la **capacité des chercheurs à identifier leurs besoins en termes de données** pour éclairer un phénomène social, économique, politique ou culturel. Le risque **d'asymétries d'information** entre chercheurs et plateformes est élevé et un **accompagnement du projet scientifique par un comité extérieur ou un régulateur** pourrait être un moyen de faciliter l'élaboration des protocoles de recherche.

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 ?

Il faudrait créer un type de registre qui liste les types de données disponibles (+ éventuellement le format dans lequel elles sont accessibles ou des exemples afin que l'on puisse voir « à quoi ces données ressemblent »). Lors de la demande d'accès aux données, un formulaire listerait les types de données disponibles et le chercheur cocherait les cases pour les catégories auxquelles il souhaite accéder.

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.

Cette tâche doit revenir au chercheur, puisque son projet est déjà validé par son institution de rattachement et/ou celles qui le financent (université, école doctorale, ANR, projets européens). Il ne nous semble pas pertinent de limiter le type de projet et les champs de recherche éligibles.

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 ?

Un simple formulaire en ligne où le chercheur décrit son projet de recherche paraît être la meilleure solution (un peu à la manière dont se déroule l'accès chercheur à Twitter actuellement). Ce formulaire serait à remplir avec une adresse mail institutionnelle afin qu'une vérification de l'affiliation puisse avoir lieu.

L'idée d'avoir des appels à projets pour encourager des travaux sur des thématiques jugées particulièrement stratégiques/d'actualité est intéressante, surtout si elle constitue une source de financement supplémentaire pour conduire ce type de recherches (une bourse par exemple).

En revanche, il serait contre-productif de conditionner l'accès aux données à ces projets de recherche « sur appel » car cela reviendrait à se priver de quantité de travaux d'intérêt général. C'est au chercheur qu'il convient d'identifier les thématiques sur lesquelles il veut travailler, et non aux plateformes (situation actuelle, dans certains cas aux États-Unis) ou à une agence nationale ou européenne, même si ces acteurs peuvent bien sûr chercher à favoriser l'intérêt des chercheurs pour certains sujets.

➤ *Evaluation des demandes d'accès et critères d'attribution :*

Les questions de cette section partent du postulat que les projets de recherche nécessitant un accès à des données de plateformes en ligne ont été définis dans le cadre de demandes formalisées (auprès d'un tiers de confiance par exemple). La question de l'évaluation de leur **qualité scientifique** se pose. Le **caractère plus ou moins innovant des projets et leur niveau de contribution à la littérature scientifique** sont des dimensions qui pourraient influencer les modalités d'ouverture des données. L'examen des demandes à l'aune de ces enjeux impliquerait **l'intermédiation de comités d'experts indépendants** pour évaluer les requêtes, selon un protocole clair et des critères transparents. Ces derniers pourraient prendre des formes différentes selon les disciplines, **tout en restant dans un cadre théorique d'habilitation préalablement défini.**

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-ils 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 ?

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** ?

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 ?

Pour vérifier le statut de chercheur, on peut imaginer une obligation d'inscription dans une école doctorale pour les doctorants, et d'être rattaché à un laboratoire pour les chercheurs.

Par contre, encore une fois il serait contre-productif de conditionner l'accès aux données à des critères similaires à ceux des financements européens, car cela exclurait de fait un grand nombre de projets qui peuvent aussi être d'intérêt public.

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.

➤ *Production et valorisation scientifique :*

Afin d'éclairer le débat public, les projets de recherche qui auront recours à l'exploitation de données de plateformes pour répondre à des questions scientifiques ont pour visée d'être **publiés dans des revues scientifiques**. Si les comités d'attribution et les plateformes ne doivent pas interférer dans les résultats et conclusions tirés par les chercheurs afin de **garantir leur indépendance**, la valorisation des travaux pourrait être reconnue, via par exemple des **protocoles de certification**. Ces derniers visent à confirmer que l'utilisation des données a été conforme au cadre réglementaire en vigueur, par exemple sur le modèle de la certification *casca*d du Centre d'Accès Sécurisé aux Données (CASD)¹⁴.

De plus, les critères de publication en sciences sociales évoluent notamment en ce qui concerne les études quantitatives et intègrent davantage aujourd'hui le **principe dit de répliquabilité des résultats** par d'autres chercheurs. Dans ce cadre, les protocoles d'analyse ayant mené à des résultats particuliers doivent pouvoir être **étudiés, critiqués, ou servir de base à d'autres travaux**. Ce principe suppose la mise à disposition des données et des ressources (codes, scripts, etc.) et peut soulever des difficultés particulières dans le cas des données sensibles collectées sur les plateformes en ligne.

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 semble plus pertinent que le chercheur ait une accréditation qui vaudrait comme preuve qu'il est au courant du droit et des bonnes pratiques dans la gestion des données (voir réponse à la question ci-dessous).

Les publications académiques sont généralement issues d'un processus de *peer review*. C'est à ce niveau là que se fait la « certification » de la méthodologie et de la 'scientificité' du papier en question.

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** ?

Les chercheurs devraient être formés en amont aux questions juridiques et éthiques relatives aux données qu'ils traitent. On pourrait par exemple imaginer que chaque chercheur doive, pour obtenir une accréditation lui permettant d'accéder aux données, valider un MOOC qui couvrirait à la fois les questions relatives au traitement de données et à la publication de ses résultats : dans quelle mesure des données peuvent-elles réellement être anonymisées ? Comment traiter des groupes fermés et autres espaces semi-privés, sur Facebook par exemple, qui sont pourtant accessibles sur demande ? Toutes ces questions mériteraient d'être clairement abordées de sorte à ce que les

¹⁴ Le CASD est un dispositif d'accès à des données sécurisées notamment d'administrations françaises (INSEE, ministères, etc.) via la mise à disposition d'un boîtier « SD-box » à des parties impliquées dans un projet d'étude préalablement agréées (universités, autorités, etc.). La certification casca-d-CASD permet aux chercheurs de signaler auprès de leurs pairs le caractère reproductible de leur recherche sur des données confidentielles hébergées au CASD.

chercheurs reçoivent tous les mêmes informations sur ces sujets - ce qui n'est pas le cas aujourd'hui.

L'accréditation pourrait éventuellement prendre la forme d'un « contrat » par lequel le chercheur s'engage à respecter les principes de « bon usage » des données tel qu'il l'aura appris pendant sa formation.

NB : Pour préserver l'indépendance de la recherche, il semble important que les plateformes n'aient pas de droit de regard ou de refus sur la publication de travaux issus de leurs données (même si elles conservent bien sûr un droit de réponse).

C. Construction des projets scientifiques: remarques complémentaires

D. Protection des données et considérations techniques

➤ *Identification des données pertinentes et construction des matériaux :*

Le terme de « données » peut recouvrir un champ très vaste (contenus, utilisateurs, archives, etc.). Délimiter son cadre d'application est donc un réel enjeu pour assurer une **cohérence entre sujets d'étude et caractéristiques évolutives des plateformes**. De plus, chaque question de recherche originale peut requérir une mise en forme particulière des bases de données d'études afin de correspondre à une méthodologie d'analyse. Par exemple, le degré de **granularité des variables**, la **composition de certains agrégats**, la **possibilité d'appareiller les données avec des bases complémentaires** issues d'autres sources sont à prendre en considération pour éviter les écueils d'un **modèle « one-size-fits-all »** qui ne permettrait pas de traiter certaines questions sous certains prismes.

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 ?
- 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 ?

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'Insee¹⁵ ?

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 peut imaginer qu'après chaque projet, les chercheurs remplissent un formulaire de *feedback* qui permettrait d'améliorer le dispositif et de le garder à jour. Ces retours d'expériences pourraient servir de base de discussion à des réunions annuelles des parties prenantes pour réévaluer le cadre d'accès.

➤ *Modalités d'accès et stockage :*

À la formulation de demandes d'accès à des données s'ajoutent des **considérations techniques** relatives aux modalités d'accès et à leur mise en œuvre. En effet, les dispositifs de mise à disposition et de partage de ces ressources doivent **être sécurisés et fiables**. Des modèles d'accès à des données via des boîtiers sécurisés ont déjà été expérimentés par des producteurs de données comme l'Insee. D'autres **modes d'accès et de stockage de ces données** pourraient s'envisager.

- 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 ?
- 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** ?
- 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** ?
- 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 ?
- 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 ?

Dans l'idéal, il serait utile d'avoir pour chaque projet une personne ressource à qui poser des questions spécifiques sur les données disponibles, les outils de consultation et éventuellement de modélisation.

Il serait aussi utile d'avoir accès à un annuaire des chercheurs/équipes ayant utilisé ces données dans leurs travaux (contact, sujet de recherche, données utilisées), afin que l'on puisse entrer en contact avec eux pour échanger, poser des questions, etc. Pour favoriser les collaborations académiques, par exemple entre chercheurs en sciences sociales et en informatique, on peut aussi imaginer des conférences/colloques ponctuels rassemblant les chercheurs ayant travaillé avec ces données afin qu'ils puissent échanger et recevoir des retours sur leurs travaux.

D. Protection des données et considérations techniques : remarques complémentaires

E. Faisabilité de l'accès et incitations

➤ *Accompagnement des chercheurs :*

La construction de projets de recherche basés sur l'utilisation de données des plateformes soulève un certain nombre de **risques relatifs aux inégalités entre disciplines ou équipes de recherche**. En effet, certaines peuvent ne pas être en mesure de proposer des protocoles d'analyse du fait de ressources limitées (capacités techniques, personnel, etc.). De plus, **le manque de connaissance des protocoles d'accès** pourrait avoir un effet **dissuasif** pour de plus petits acteurs, par exemple moins bien financés ou moins en capacité de répondre à des appels d'offre nationaux ou européens.

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 ?

Dans l'hypothèse d'un comité de sélection indépendant des projets, il est important que les décisions rendues soient suffisamment détaillées pour que les demandeurs ayant été refusés pour des raisons de non-conformité puissent comprendre la décision et éventuellement revoir le projet s'il peut être amendé.

Il serait aussi utile d'avoir une formation des chercheurs aux outils de consultation, et surtout au statut légal des données pour que nous soyons au clair sur ce que l'on a le droit de faire/de publier avec ou non. Voir notre réponse à la question C.9 sur l'accréditation.

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 ?

- Ne pas restreindre l'accès à ces données à de gros projets européens interdisciplinaires et/ou impliquant un certain nombre d'institutions.
- S'il faut se déplacer pour accéder aux données, une possibilité de financement du trajet serait intéressante.

➤ *Incitations des plateformes :*

L'accès des chercheurs aux données des plateformes en ligne vise à améliorer la compréhension des dynamiques socio-économiques, politiques, culturelles et de fait, **pourrait justifier la participation des plateformes dans le cadre par exemple d'un dispositif de contribution à la connaissance scientifique**. Elles pourraient également bénéficier des résultats des recherches menées, ce qui contribuerait à faciliter leur dialogue avec le monde de la recherche.

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 ?

En aucun cas des incitations seules pourront contraindre les plateformes à faire les investissements nécessaires pour ouvrir leurs données. Un cadre légal doit être imposé, sans quoi leurs choix suivront leurs propres intérêts, tuant dès lors toutes études à leur propos. C'est à la recherche, aux journalistes, ou aux associations de définir les données dont elles ont besoin pour mener à bien leurs missions.

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 ?

Oui, l'exemple du CESP paraît correspondre aux besoins. Son rôle réside dans l'analyse des décisions rendues, et surtout dans les raisons qui conduisent à un refus.

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

Oui, cet aspect est nécessaire sans quoi l'ensemble du dispositif ne pourrait pas exister. Les plateformes vont opposer des résistances, parfois non explicites, ou cachés, dans les données transmises, le temps de réponse, les manquements, etc. Il faut que ce comité d'audit puisse les noter et les signaler.

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

Dans un premier temps, il peut être envisageable de sortir du cadre d'étude les données liées à la publicité ou aux annonceurs. Ces données sont sensibles pour les entreprises, les prendre en compte risque de bloquer l'avancement de recherches qui portent sur des sujets déconnectés des enjeux financiers. L'intérêt du grand public et de la recherche ne réside pas uniquement dans la connaissance des revenus générés par ces plateformes.

E. Faisabilité de l'accès et incitations: remarques complémentaires
