Arcom

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

INRIA – Projet Pilote REGALIA

Arcom

Le régulateur de la communication audiovisuelle et numérique

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.

[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.

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.

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 1er 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 transparence[2]. 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 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 concernées[4].

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

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

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].

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 recherche[7]. 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 évolutions[9]. **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 ligne[10], des obligations

[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)

[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=Th

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 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

e%20DSA%20proposal%20is%20a%20horizontal%20instrument%20putting,and%20Digital%20services%20act%20%28DSA%29%20draft%20asymmetric%20obligations

[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

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 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 publics[12].

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

[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.

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 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.) ?

Oui, nous (projet Regalia, Inria) avons mené des recherches utilisant des données issues de plusieurs plateformes en ligne (au sens DMA du terme) :

- Des plateformes de livraison de repas à domicile (de type Uber Eats ou Deliveroo)
- Des plateformes d'agence de voyage en ligne (de type Booking.com ou Trivago)
- Des plateformes de location de biens de courte durée (de type AirBnB ou Abritel)

Nous les avons collectées par « scraping ».

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

Trois types de difficultés rencontrées :

- Techniques : éviter que l'outil de scraping soit facilement identifié comme tel par la plateforme
- Commerciales : les variations des offres commerciales de la plateforme rendent difficiles la pratique de sondages stables pour mesurer des effets algorithmiques
- Sémantiques : les algorithmes à observer utilisent parfois des variables cachées, non accessibles par le web, qu'il est difficile d'approximer ou de synthétiser

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.

En partenariat avec le PEREN, nous avons dû renoncer à l'étude de l'algorithme de recommandation d'une plateforme de contenu (de type réseau social) par refus d'accès (accès non conforme aux CGUs).

En partenariat avec le PEREN, deux projets d'études en collaboration avec deux grandes plateformes sont ralentis par la difficulté à poser le cadre de la donnée collectée. Du fait des deux plateformes. Ralentis = plusieurs semestres.

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 des cas nous avons informé la plateforme des études en cours, c'est probablement un facteur de succès comparé aux approches collaboratives d'après notre expérience. Il est probable que dans chacune des études algorithmiques que nous avons menées à bien, deux critères de succès en ont permis la réalisation :

- Le fait que nos scraping concernaient un tout petit volume, étalé dans le temps, des requêtes possibles à la plateforme. Nous avons favorisé des comptes sans historiques (cold user) pour éviter la création de faux comptes ou la vampirisation de vrais comptes par nos bots
- Les plateformes étaient des plateformes commerciales et non des réseaux sociaux (elles ont dont une motivation économique à laisser accès à leurs données par des bots)
- Il est possible que les volumes extraits et le caractère ponctuel dans le temps et dans l'espace des requêtes de nos recherches soient passés « sous le radar » des détecteurs de bots des plateformes. Ceci est une hypothèse.

➢ *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 ? Partenariat scientifique avec le PEREN depuis Mars 2021
- concerne-t-elle une ou plusieurs **catégories de bénéficiaires**

particuliers (chercheurs, ONGs, entreprises, etc.) ? Une : Bercy

- existe-il des **critères de sélection** de ces bénéficiaires ? Si oui, lesquels ? Nous sommes plutôt les sélectionnés que les sélectionnants, en l'espèce
- quel(s) **type(s) de données** cette politique concerne-t-elle ? Données produits et commerciales (repas, restaurtants, livraison, hôtels, bed & breakfast)
- intègre-t-elle **un volet de contrôle ou de suivi** de l'utilisation qui est faite des données délivrées ? Oui, par comité de pilotage conjoint des deux entités + démarche RGPD (risques faibles en contexte de données commerciales ici).

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.

L'équipe Regalia est jeune (Juin 2020) et nous n'avons pas encore initié de collaboration par la data avec d'autres équipes de recherche. Nous le ferons probablement en 2023, grâce au lancement de doctorat co-encadrés avec une ou deux équipes de recherche Inria.

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

Regalia n'observe pas le contenu en tant que tel des données des plateformes, mais principalement les effets des algorithmes de Recommandation ou de Pricing des plateformes. Par exemple, sur une plateforme audiovisuelle, nous pourrions regarder si des contenus « toxiques » (tel que définis ou taggés par une autorité) ne sont pas plus poussés vers un client de catégorie A que vers un client de catégorie B. Pour ce faire nous n'avons pas forcément besoin de collecter des data de client de type A ou B, mais nous pouvons les générer « fictivement » ou pseudo-manuellement. A l'image des travaux réalisés par le CSA, manuellement, ou par l'équipe Inria/Wide, robotiquement.

Nous sommes donc relativement atypiques dans notre besoin de collecte, puisque focalisés sur la réponse algorithmique de la plateforme auditée et non sur son contenu. Bien entendu, une part du catalogue « produit » de la plateforme est collectée, ce faisant, qui peut soulever des questions pour la plateforme.

D'autre part, le fait d'utiliser intensivement, même localement, des requêtes générées par des robots peut perturber la plateforme sous trois angles :

- Créer un déni de service local
- Fausser les algorithmes de la plateforme (regarder des hôtels sans les acheter donne l'illusion à la plateforme que les prix sont trop élevées, par exemple)
- Potentiellement, requêter une plateforme d'une manière systématique, même ponctuelle peut permettre d'effectuer un « reverse engineering » de l'algorithme de la plateforme, et donc venir dérober sa propriété intellectuelle.

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 réflexion[13]. 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) ?

Avant de définir les critères de limitations, je poserais volontiers les risques que l'on cherche à couvrir. J'ai du mal à les cerner à ce stade, mais je pense à trois formes de risques :

1. Des acteurs peu scrupuleux ou à la déontologie légère, se servent des données des plateformes pour faire le buzz sur des thèmes sensibles. Ex : un acteur d'extrême-droite qui montrerait que des contenus de personnes issus de l'immigration sont plus violents ou sexistes que de personnes de souche. Un acteur pro-Woke qui pillerait toutes les photos d'ouverture d'huîtres sur facebook pour prouver que l'ouverture d'huître est une activité genrée.
   Cela a déjà été le cas lors du covid donc on peut imaginer une démultiplication de fakenews si n'importe quel groupe politique (ou avec un faux nez d'ONG) peut exploiter des données riches des plateformes.
2. Trop d'acteurs malveillants extraient des données de la plateforme ou viennent perturber les algorithmes en faussant les comportements clients. C'est une variante algorithmique du déni de service.
3. Des acteurs économiques concurrents (ou dans l'éco système) qui viendraient piller les données pour en faire un usage commercial, voire qui viendrait piller l'algorithme pour le reproduire ou en déduire des stratégies frauduleuses. Exemple si j'audite bien l'algo détecteur de faux comptes de Deezer, je peux le contourner et doper les audiences, donc le revenu, de mon rapeur préféré.

Il me semble important de couvrir ces risques. A priori, la déontologie journalistique ou scientifique devrait suffire. Mais pour les ONG, les thinktanks ou les associations de consommateurs je pense qu'une forme d'enregistrement / déclaration est nécessaire, probablement accompagné par une forme de supervision.

[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).

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

Un raisonnement basé sur les risques, tel que celui évoqué, et sur le moyen de mitiger/couvrir ces risques, par catégorie d'acteurs, devrait pouvoir amener à des possibilités d'accès (en volume, en profondeur, ou en supervision, par exemple) différentiés. Ceci afin de ne pas brider des acteurs « à faible risque » dans leurs études et d'apporter un niveau de surveillance raisonnable aux autres.

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* ?

A l'image de ce que les villes ou l'état fournissent à ce jour en Open Data ou exigent de certains de leurs prestataires utilisant des ressources qui leur sont mises à disposition, il me semblerait souhaitable que tout acteur numérique ayant un impact systémique (maîtrisant et ayant une vision représentative des données d'un espace numérique et de ses usages) fournissent un ensemble de données défini par exemple par une instance de régulation.

Toutefois, le protocole d'anonymisation devra être soigneusement étudié pour évier des dérives inévitables (comme chacun sait, le croisement de quelques variables : genre, date de naissance, code postal, suffit à connaître individuellement 80% des individus).

- *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 ?

La réponse ne peut être idéologique. Il faut se concentrer sur l'efficacité du tiers de confiance et de son périmètre culturel. Exemple : une autorité de régulation n'ayant pas de culture algorithmique des outils de recommandation sera inadaptée pour surveiller les dérives d'un outil de recommandation à la pointe de l'état de l'art.

L'autorité la mieux à même de jouer ce rôle de tiers sera celle :

- Qui dispose d'une véritable culture data et algorithmes, avec dans son board des experts techniques de ces sujets, ayant eu des parcours à la fois académiques et industriels sur ces technologies
- Qui dispose d'une masse critique suffisante pour la création d'une telle équipe
- Qui ai la capacité politique (pour peser sur les plateformes) et judiciaire à agir, avec une vélocité avérée à la faire (fameux contre-exemple de la CNIL irlandaise)
- Qui dispose d'une sensibilité culturelle locale (sociologique, politique) lui permettant de comprendre les enjeux et les comportements clients sous-jacents afin de bien cibler les datas à collecter et les « featurer » (les pré-traiter) avec la bonne sémantique initiale.

Le protocole « tiers auditeur » que j'imagine serait spécialisé par type de risque puis type de plateformes (voire typologie d'algorithme à surveiller dans les plateformes quand l'algorithme est le problème).

On pourrait penser à :

- Par type de risque/plateforme : Un protocole de collaboration entre le tiers auditeur et des instances de conformité (sur le format bancaire) de la plateforme. Sans cette collaboration, les audits et récoltes de données seront fortement limités. Les risques de blocage ou barrières sont avérés.
- Des protocoles itératifs avec des jalons à 6 mois pour juger de la bonne volonté des plateformes et de la performance (ou des verrous techniques rencontrés) du tiers auditeur. La possibilité de disqualifier un tiers auditeur mal organisé doit être rendue possible
- La mise en place d'une structure d'hébergement (entrepôt d'audit) de type entrepôt des données de santé, sur un espace souverain. L'historisation des données pourra permettre d'effectuer des études différentielles et longitudinales. C'est à ce niveau également que des anonymisations profondes pourront être menées à bien. (cf APHP)
- La création de « couches » d'analyse faisant intervenir des types d'acteurs différents. Par exemple, des architectes, des data engineers, des data scientists, mais également des chercheurs en science sociale ou psychologues (étude de la dépression, étude des biais de genre, étude de la viralité des contenus toxiques, ou des expertises artistiques le cas échéant).

Selon le protocole d'accord avec la plateforme, certains sujets d'étude pourront faire intervenir des équipes de R&D de la plateforme (ex : AirBnB et les biais raciaux dans la location de biens).

- A l'image des travaux réalisés sur l'entrepôt des données de santé de l'APHP il est illusoire de fonctionner séquentiellement selon la métaphore de l'extraction pétrolière (creuser, extraire, raffiner). Les équipes des différentes couches impliquées doivent pouvoir interagir pour modifier les agendas et les priorités des autres couches, par exemple en relâchant certaines contraintes ou en faisant évoluer le périmètre des données collectées au fil de l'eau.
- Il est également inopérant de fonctionner exclusivement sur « commande » c'est-à-dire de n'extraire la donnée qu'une fois qu'elle a été demandée par le partenaire. Non seulement la demande a de grandes chances d'être mal formulée (car le partenaire ne sait pas de quoi la donnée est faite) mais le temps de latence de la réponse sera prohibitif si le tiers auditeur « découvre » la donnée à la demande.
- Le facteur d'échec numéro un sera la difficile et lente montée en compétence technique du tiers auditeur. Peu d'autorités de régulation ont à ce jour une équipe de 150 personnes en ordre de bataille pour collecter et analyser des données d'une variété de plateformes. Un second facteur sera probablement l'inflation des demandes de collectes, sous perfusion politique des donneurs d'ordre (scientifiques ou régulateurs) en présence. On retrouve ce type d'échecs dans la création des data-labs de grandes entreprises du CAC40 ou équivalent en Europe. Il y a peu de raisons qu'un tiers auditeur puissant hébergé dans un service régalien ou une autorité indépendante fasse mieux. Le taux d'échec des projets de collecte et traitement de données de grande taille, même au sein d'un groupe propriétaire de ses données, est très élevé (75% dans ma propre expérience, en France et en Europe).

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 ?

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 ?

Dans mon idée le « demandeur » et le « tiers auditeur » n'auraient pas une relation client-fournisseur. Dans le sens où le premier commanderait une prestation d'extraction (type extraction pétrolière) à l'autre. Les risques de mauvaise extraction ou de désalignement seraient de mon point de vue

trop élevés et aboutiraient à des effets tunnels non performants. Pour continuer dans la métaphore pétrolière, il s'agit plutôt de puits d'exploration et de « paris » successifs pour identifier les formes du réservoir.

Le demandeur fournirait des ressources (scientifiques ou techniques) dans les couches « métier » (sociologiques, psychologiques, économiques) en association avec les équipes du tiers auditeur. Une gouvernance conjointe et agile (le périmètre pourra évoluer en quelques mois au fur et à mesure des découvertes ou des déceptions) s'assurera que :

- la demande est bien formulée. Si la demande évolue au fil des découvertes ou des blocages constatés, une reformulation devra être proposée voire l'abandon de la demande
- la somme des moyens alloués dans les différentes couches couvre le besoin de manière prévisionnelle
- cette agilité est cruciale dans la réussite du montage de la collaboration inter-couches

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 ?

Comme dans le secteur privé, la non performance du couple « demandeur/tiers auditeur » aura un impact sur le portfolio des projets en cours. Des cycles très courts d'état d'avancement, à un niveau très technique, et centrés sur la demande et ses besoins en données, devraient permettre de nettoyer le portefeuille et de donner plus de ressources aux projets bien calibrés. L'existence de « product owner » ou « demand owner » de qualité apte à faire le pont entre une demande « métier » et les données nécessaires à extraire est un facteur de succès de toute l'approche. Une implication technico-politique (dans les « steering committee ») sera nécessaire au recadrage inéluctable des ambitions et des budgets locaux.

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

Je préfère l'efficacité à la transparence dans ce cas d'espèce. La transparence me semble ici du second ordre, compte tenu du retard pris par les autorités de régulation dans la collecte et la compréhension des process à l'oeuvre dans les grandes plateformes.

Mais, une vision claire du « portfolio » des demandes, ce qu'on en attend, les partenaires, les itérations, et les freins, est un prix à payer pour s'assurer d'éviter les redondances et les effets discrétionnaires éventuels. Une vision budgétaire (tant de demandes pour les ONG, tant de demandes pour les associations de consommateurs, pour les chercheurs en contenu haineux, en anoréxie etc …) pourra à grosse maille permettre aux décideurs de dimensionner convenablement les efforts dans les différentes couches. Pour la partie recherche, une forme de budgétisation a priori (type plan IA) sur

des axes stratégiques, évitera la perte de temps administrative bien connue des réponses à projet (de type ANR).

Le « tiers auditeur » devra pouvoir influer sur les priorisations des projets pour atteindre des objectifs de performance afin d'éviter les effets tunnels des projets magnifiques et richement dotés mais visibles à 3 ans.

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

Si les conditions qu'elles posent à leur participation sont raisonnables (en termes de publication et de contrôle de la propriété industrielle) elles sont un des acteurs parmi d'autres et peuvent même co-financer, à travers des labs conjoints avec le tiers-auditeur, des demandes (ou nuages de demandes) communes avec des équipes de recherche. La donnée collectée fera partie de l'asset (entrepôt) du tiers auditeur.

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

Pour avoir participé à la création d'une équipe de 150 personnes (architectes, data-scientists, data-engineers, UX, Product Owner) au sein d'un grand groupe international (40 milliards de CA), je suis conscient des risques inhérents à la création de grands projets data technico-politiques. Le taux d'échec des projets est considérable et les frustrations non moins significatives. Nous n'en avons pas les moyens.

Le premier risque est culturel. Les cadres actuels des autorités de régulation, tout comme les cadres des grands groupes industriels comme ceux que je connais, sont peu formés à piloter de tels projets, au-delà de la recherche du budget de lancement, de la recherche de locaux et de la journée d'inauguration ☺.

Le second est humain : recruter quelques dizaines de personnels techniques sur ces sujets, avec les contraintes budgétaires de l'état (ou d'une autorité indépendante) est quasiment impossible. Or le profil de ces techniciens (ou des Product Owners) est proche de ceux utilisés par les grandes plateformes elles-mêmes qui ont mis la barre très haut en terme de rémunération et de cadre de travail. Même Total, Danone, Carrefour ont eu de gros ratés dans le lancement de leurs équipes data, avec des retards de plusieurs années et pourtant des enjeux forts et portés par la direction générale.

Le troisième est technique : explorer des données massives en tâtonnant pour identifier des « patterns » ou répondre à des questions métiers est plus compliqué que de lancer un nouveau logiciel de paye. En effet :

- les plateformes ont 10 à 20 ans de culture data, avec des bases gigantesques et des pipeline de données en perpétuelle évolution. Faire des extractions, même en cas de collaboration, amènera de lourdes incompréhensions culturelles, voire des effets de

dissimulation, plus ou moins intentionnelle

- à l'image de l'entrepôt des données de santé (type APHP), il est très difficile de faire dialoguer des experts transdisciplinaires (exemple : des experts de l'anorexie, des data scientists, des data engineers, des spécialises du traitement d'image ou du langage, si l'on veut mesurer l'effet d'Instagram sur l'anorexie des adolescentes). Monter une équipe conjointe, au point de collecter le bon échantillon, et valider des hypothèses, demandera des efforts véritables d'ouverture et d'adaptation. Sans compter le rapport au temps qui n'est pas le même selon les disciplines.
- Enfin, si la partie noble est souvent vue comme « la publication », les faces cachées de la captation de données, leur nettoyage, mise au propre, anonymisation, sont souvent les parents pauvres et la pression à baisser la qualité pour augmenter le débit de sortie des études, nécessitera une fine gouvernance. Eviter les « dettes techniques » sera un enjeu particulièrement délicat.

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 me semble très difficile de juger la « pertinence » d'une demande sur la seule base du pedigree des chercheurs. Et pourtant, pragmatiquement, une équipe de recherche qui a accumulé 15 ans de travail sur les contenus haineux devrait a priori tirer parti plus rapidement d'un accès à Twitter par exemple.

Peut-être serait-il possible de créer plusieurs catégories de demandeurs. Chercheurs à pedigree et « jeunes » chercheurs. Pour appliquer aux premiers des règles classiques de pertinence et aux seconds des critères différents, favorisant l'innovation et la création d'expertise.

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.

1. On peut imaginer quelques grandes missions d'intérêt général dans le cadre desquelles il soit plus facile de candidater et d'être accrédité (haine en ligne, fakemed, etc …). Pour lesquelles en plus d'autorisations on peut ajouter des financements type ANR.
2. Mais il faut également des thèmes innovants, hors d'un cadre pré-établi. Si je pense à la censure des photos de femmes obèses sur Instagram, l'idée n'est peut-être pas née dans un ministère, mais par constat dans la société civile (constatant une discrimination par exemple, de la part du groupe discriminé), ou dans un groupe de chercheurs/ONG.
3. Une courroie de transmission avec certaines entités « lanceuses d'alerte », de type association de consommateurs, conspiracy watch etc … pourrait être un canal d'inspiration

Les trois formats doivent co-exister, le premier pouvant être sponsorisé pour favoriser la variété et la multiplicité des approches.

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 ?

Pour le mode 1 (missions d'intérêt général), l'appel à projet le moins administratif possible doit être favorisé, et la création d'une communauté d'intérêt avec quelques

exigences de communication/partage pour favoriser la coordination.

Pour le mode 2 et 3, en auto-saisine en quelque sorte, je crois qu'il faut définir un bureau de requêtes qui puisse trier les demandes et les prioriser si elles ont un impact fort sur les ressources de collecte. C'est une tâche lourde mais du même ordre qu'etalab par exemple, qui reçoit des demandes et arbitre l'utilisation de ses ressources. L'arbitrage pourra être assez multipartite pour être certain que des points de vues minoritaires puissent y être représentés. Des fiches projets assez structurées devraient permettre d'accélérer le filtrage et permettre éventuellement du dialogue avec le comité de sélection.

- *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-il y avoir une place et un rôle et, si oui, lequel ?

Absolument pertinent.

Y sera effectué un suivi des rôles de l'équipe demandeuse (chercheur ou ONG) mais également de la prestation du tiers auditeur pour que le système s'améliore et que les futurs arbitrages et filtres soient plus efficaces.

Pour les projets « recherche » : comité scientifique + représentant du régulateur concerné (ex : si RGPD alors CNIL) + acteur du monde technologique (issu du lobby des plateformes) + services régaliens si sujet stratégique (mission d'Intérêt Général)

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** ?

Avis personnel : sur tout un ensemble de sujets, la plateforme n'a pas à être présente, soit parce que son rôle n'est pas utile, soit parce que l'on ne souhaite pas qu'elle soit informée des résultats.

Mais il semble souhaitable de permettre à la plateforme de collaborer, de manière officielle et avec de vrais moyens (évalués a posteriori) à certaines études. A la fois pour favoriser les échanges (exemple : la plateforme a déjà des outils de modération qui détecte des contenus haineux, il peut être utile de s'en servir pour détecter des contenus haineux ET antisémites par exemple) mais aussi éviter des erreurs d'analyses, les équipes du tiers auditeur ou les chercheurs pouvant mal interpréter des résultats.

Permettre des collaborations est à long terme vertueux car crée une percolation d'idées et de connaissance entre les équipes.

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 ?

Inter disciplinaire et européen, avec des compétences techniques mesurables (pour éviter la surcharge du tiers auditeur).

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 ?

A l'image des équipes projet (type Inria) qui renouvellent tous les 4 ans leur existence et missions, l'aspect temporel est critique, quitte à permettre et à valoriser des changements de points de vue en court de projet, au fur et à mesure que la data se collecte et permet de tirer des analyses.

- *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 *cascad* 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**

[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 boitier « SD-box » à des parties impliquées dans un projet d'étude préalablement agréées (universités, autorités, etc.). La certification cascad-CASD permet aux chercheurs de signaler auprès de leurs pairs le caractère reproductible de leur recherche sur des données confidentielles hébergées au CASD.

***réplicabilité* 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 ?

Une forme de certification externe semble nécessaire selon au moins trois critères : reproductibilité, patrimonialisation (façon open data par ex.) et anonymisation (en plus de RGPD bien entendu).

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** ?

Il n'y a pas de différence entre une étude faite avec des données issues de Twitter (récoltées par le tiers auditeur) et une étude faite par des chercheurs de AirBnB sur des données AirBnB (exemple : discrimination des afro-américains aux US dans l'accès à la location). Le cas doit être déjà traité par les laboratoires de recherche, tout comme c'est le cas en médecine. Un laboratoire de recherche en socio, doit avoir une déontologie et des process de validation pré-publication qui couvrent ces cas.

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 ?

Le secteur privé est déjà dans ce cas quand il stocke des données clients ou des données capteurs pour des besoins génériques ou spécifiques partiellement connus, il n'y a rien de particulier ici. Des plateformes logicielles (ex : Palantir ou Dataiku) existent qui permettent de modéliser des flux de traitement successif des données, en les « décorant » des actions menées (nettoyage, enrichissement, complémentation, anonymisation, « featurisation ») ainsi que des auteurs de ces actions. On peut imaginer la création de plateformes souveraines sur ce sujet (il doit en exister, même partielles)

Cela permet un équilibre entre a) ne pas dupliquer toutes les sources de données à chaque besoin spécifique b) être capable de spécialiser la donnée pour la réponse à certaines questions afin d'être performant et frugal.

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 ?

Ceci me semble un dommage collatéral heureux mais pas un objectif en soi. Le chercheur obéit à une logique extrêmement spécialisée dans un cadre très fermé et cadré. Alors que des indicateurs sont plus souvent exprimés en langage métier, voire grand public avec des objectifs de surveillance ou de contrôle. L'un peut nourrir l'autre, mais le processus de distillation est lent et hasardeux (part des anges élevée ☺).

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

Ne serait-ce que par paresse et percolation des idées oui. Trop insister sur cet objectif peut être toutefois opposé à la logique d'agilité et d'efficacité et donc

générateur d'usine à gaz au nom de la mutualisation collaborative. La priorité est au débit de réalisation de demandes. Mais petit à petit, des blocs de données seront populaires et naturellement les chercheurs d'un labo B auront tendance à partir des données du labo A, quitte à les enrichir, pour prouver un résultat additionnel. Egalement, l'angle du régulateur, qui ajoutera d'autres données (Insee, ou météofrance) pourra donner une valeur supplémentaire au data-set.

Idéalement, il faudrait que la richesse des données collectées soit facile à appréhender (naviguer, rechercher, filtrer), mais je me doute que c'est un procédé très délicat compte tenu de la largeur du spectre. C'est un challenge considérable.

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 ?

Certains cadres seront assez stables (par exemple la notion de « mur » sur Facebook) pour que la gouvernance et le mode de stockage/extraction n'évolue pas trop vite. Mais les véritables innovations (exemple le « Pour toi » sur TikTok, qui capture les quelques secondes de réaction de l'internaute) demandera de remettre en cause les collectes et certains arbitrages.

Si l'on fait le parallèle avec Google Analytics, qui capture toutes les données web d'un site. Je suppose que le produit ne change pas radicalement tous les ans même si des évolutions profondes (RGPD) le traversent de temps en temps.

- *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 ?

Je pense à trois grands modes :

1. Le scraping web
2. L'accès à une API
3. La demande explicite de données massives à la plateforme
4. Et une hybridation de 1 ou 2 avec 3 pour récolter des données complémentaires au fil de l'analyse

Le scraping (à condition qu'il soit toléré par la plateforme, ce que l'on considèrera comme acquis) a l'avantage d'interroger la plateforme dans ses conditions réelles d'usage. Il suppose toutefois de ne pas être détecté par la plateforme comme usage artificiel (sinon la plateforme se comportera artificiellement) et de pouvoir générer tout le flux de data nécessaire à la « requête » web réalisée. Ceci peut demander une

certaine forme de collaboration de la plateforme. D'où le point 3, ou certaines données sont fournies par la plateforme en complément du scraping.

Il a l'inconvénient d'être sensible à l'état algorithmique de la plateforme, qui peut être mouvant, et donc les analyses seront « datées » et devront être renouvelées à intervalle régulier pour être pertinentes.

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** ?

Mêmes standards que les données de santés, avec des chercheurs/régulateurs disposant de droits d'accès spécifiques, sur déclaration de projet.

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

Mêmes exigences techniques que les données de santé

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 ?

Evaluation des risques initiaux (à la déclaration d'intention du projet), et suivi des procédures de protection. Bilans par acteur et nomination de responsables « protection de données » dans les équipes de recherche ou de régulation, pour ces projets spécifiques.

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 ?

Le soutien technique semble nécessaire pour pouvoir a) s'assurer d'une bonne compréhension de la donnée collectée par rapport aux enjeux du projet b) mutualiser autant que faire se peut, les traitements de données réalisés c) faire bénéficier chaque projet d'une expertise pointue du tiers auditeur

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 ?

A l'image de ce qui se fait dans le secteur privé, les différentes sources de données sont regroupées dans des « data products » avec des responsables de ces produits (« product owner ») pour lesquels les chercheurs seraient en quelque sorte des clients. En accordant une valeur à ces clients (petits ou gros, compétentes techniquement ou non) le product owner peut budgétiser des travaux techniques et aller plus loin que la simple mise à disposition afin d'éviter une barrière d'entrée trop grande. Toutefois de trop petits acteurs ou manquant de compétences techniques seraient encouragés à s'associer avec des partenaires plus compétents, à la façon des projets européens, pour éviter des déperditions d'énergie de la part du tiers auditeur.

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 ?

Comme dans le secteur privé, au sein des grosses plateformes de data, une équipe de « forward deployed data engineers/data scientists » pourrait être dédiée à la création d'univers de données utilisables par des équipes trop petites ou faibles techniquement. La taille de cette équipe pourra être définie par budget au vu des choix politiques de diversité des projets (entre gros et petits, chercheurs et lanceurs d'alerte, etc.).

- *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 ?

Les plateformes auront trois incitations naturelles à participer :

- lobbyiser les équipes techniques (du tiers auditeur) et de recherche, pour faire passer la bonne parole. Cela marche déjà très bien.

- surveiller l'état de l'art des technologies/méthodologies/études créées, pour anticiper les conséquences de la mise en conformité de leurs propres pratiques et algorithmes

- disposer d'un vivier de chercheurs pour faciliter les recrutements

Ce me semble suffisant pour un acteur soucieux de sa bonne santé économique dans l'éco-système européen.

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 ?

Je suis dubitatif sur la pertinence d'un comité amont, compte tenu de l'absence d'expertise sur ces sujets à ce jour. Il risque donc de basculer dans le principe de précaution et de freiner la démarche, qui a déjà beaucoup de handicaps.

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

Absolument, basée sur les faits et les retour d'expérience des premiers projets et du tiers auditeur.

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

Si des équipes de recherche abordent des projets susceptibles de révéler un secret des affaires, il serait intéressant que la plateforme, pour chaque projet (ou axe = portefeuille de projets connexes)

- explique les conditions (de volumétrie, de couverture fonctionnelle, de périmètre de la donnée) dans lesquelles un secret pourrait être révélé ou un déni de service opéré
- motive des restrictions dans les accès par des éléments quantitatifs
- en une approche contradictoire avec l'équipe de recherche, le tiers auditeur et l'autorité de régulation concernée. Sous contrainte temporelle forte (par défaut le projet est lancé au bout de p semaines sans retour).

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