Arcom

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

Criteo

Arcom

Le régulateur de la communication audiovisuelle et numérique

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**

[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.

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

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

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 éclairages sur les questions de désinformation en ligne. Dans cette perspective, l'EDMO a au titre de ses objectifs

[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/

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 supplémentaires sont prévues afin d'augmenter encore le niveau de transparence de leur action, notamment

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

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 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

[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

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.) ?

A ce jour, Criteo n'a pas fait d'usage systématique de données issues d'une ou plusieurs plateformes en ligne (même si Criteo a déjà eu recours à des données issues de l'API publique de Twitter pour nourrir certains travaux de recherche en Natural Language Processing (Traitement Automatique du langage – TAL). Criteo fait donc un usage ponctuel des données issues des plateformes en ligne, du fait de l'accès difficile et souvent payant à ces données. Criteo soutient un accès facilité aux données des plus grandes plateformes, sous couvert de la protection de la vie privée et du secret des affaires, afin de faciliter la recherche sur des problématiques sociétales. Un accès facilité aux données des plateformes en ligne permettrait par ailleurs à Criteo de mieux comprendre les phénomènes et dynamiques de propagation de contenus préjudiciables en ligne. Cette meilleure compréhension permettrait de considérablement affiner les outils qui servent aujourd'hui à éviter que des emplacements publicitaires financent des sites diffusant des contenus préjudiciables, comme la désinformation par exemple.

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

La première difficulté relative à la collecte des données issues d'une ou plusieurs plateformes en ligne relève du caractère confidentiel de ces données et, de facto de leur absence d'accessibilité. La seconde difficulté se révèle souvent d'ordre financier - en deçà d'un certain volume, l'accès aux données demeure gratuit, mais au-delà, l'accès aux données devient payant.

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.

Peu de projets de recherche menés par Criteo reposent à l'heure actuelle sur des données issues d'une ou plusieurs plateformes en ligne. De ce fait, Criteo n'a jamais abandonné un projet du fait de l'impossibilité d'accès à ces données. Néanmoins, dans le cadre de la lutte contre la désinformation, un accès facilité aux données des plateformes en ligne nous permettrait d'observer davantage les dynamiques à l'œuvre et de participer à la réplicabilité des résultats des projets de recherche menés par lesdites plateformes.

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.

N/A

➢ *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 ?

Depuis 2014, Criteo a été pionnier dans la mise à disposition de jeux de données à la collectivité (chercheurs en machine learning, ONGs, entreprises). Criteo détient d'ailleurs le record de mise à disposition de l'un des plus grands ensembles de données au monde (accessible sur https://ailab.criteo.com/ressources/). Cette politique de partage des données Criteo a plusieurs finalités :

- *Etablir Criteo comme un acteur clef de la recherche en IA au niveau mondial ;*
- *Permettre à la recherche académique d'aborder des problématiques pertinentes pour l'industrie ;*
- *Assurer la réplicabilité par tous des recherches menées par Criteo (données ouvertes et codes open source).*

Les données ainsi partagées à des fins de recherche incluent entre autres des données d'entraînement de modèles statistiques (pour détecter les populations des

consommateurs sensibles à une offre commerciale), des prédictions sur les enchères en lignes, etc… Tous nos ensembles de données ont été anonymisés en ligne selon les normes de confidentialité et de protection de la vie privée les plus élevées et sont accessibles ci-dessous via notre site :

- *Criteo Uplift Modeling Dataset (CRITEO-UPLIFT-1)*
- *Criteo Sponsored Search Conversion Logs*
- *Criteo Attribution Modeling for Bidding Dataset*
- *Kaggle Display Advertising dataset*
- *Criteo 1TB click logs*
- *Dataset for evaluation of counterfactual algorithms*

Dans la mesure où il s'agit de données anonymisées, Criteo n'a pas mis en place de politique de sélection des bénéficiaires, ni de politique de contrôle ou de suivi de l'utilisation qui est faite de nos jeux de données publics autre que le système de licence "opensource" Creative Commons (CC-BY-SA-NC) sous lequel Criteo opère pour définir les usages autorisés.

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 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) ?

Pour un défenseur de l'Internet ouvert comme Criteo, une recherche ouverte est essentielle pour promouvoir un écosystème sain et transparent. Criteo n'est donc pas favorable à une limitation en amont des types d'acteurs bénéficiant de l'accès aux données, à partir du moment où l'entreprise a pu s'assurer en amont que les données qu'elle met à disposition ne contiennent ni secret industriel ou commercial, et ne portent pas atteinte à la vie privée.

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

_____
_____
_____
_____
_____
_____

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* ?

Il est fort probable que le grand public n'ait que peu d'intérêt pour des jeux de données de recherche anonymisées.

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

[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).

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 **tie**

i) Si oui :

- ce tiers de confiance devrait-il être un acteur public **européen ou national** ? Dans ce ca
- quelles pourraient être les **modalités d'organisation** d'un **protocole fléché et encadr**
- Les modalités d'implication du tiers de confiance seraient-elles à définir selon le **niveau d**

L'intervention d'un tiers de confiance devrait être volontaire. Une intervention obligatoire serait uniquen

ii) Si non :

- pour **quelles raisons** ? Celles-ci peuvent être diverses : juridique, académique, logistiqu
- un modèle **d'interaction direct** entre la plateforme et les chercheurs vous apparaît-il pr

_____

_____

B.4. Dans l'hypothèse d'un mode de régulation qui impliquerait l'intervention d'un **tiers de confian**

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

Nous pensons que le modèle de gouvernance du supercalculateur Jean Zay pourrait offrir une réponse ad national de calcul intensif ou supercalculateur dédié à des applications d'IA, dont l'installation et modali

ii) quels **garde-fous** pourraient être mis en place pour assurer un accès à des données perm

_____

_____

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

Si l'ouverture des données des plateformes en ligne à des fins de recherche était soumise à l'interventi requêtes refusées, ainsi que les motifs de refus.

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

_____

_____

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

_____

_____

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 ?

_____
_____
_____
_____
_____

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.

Il serait judicieux de favoriser la reproduction ou réplicabilité d'une partie des résultats des projets de recherche des plateformes concernées – par exemple, en imposant aux plateformes de publier de manière systématique les données sur lesquelles elles ont appuyé leurs projets de recherche (sous réserve que ces données ne portent pas atteinte à la vie privée des personnes et ne contiennent ni secret industriel ou commercial). Par ailleurs, les données concernées par le partage ne devrait pas être restreintes à des champs de recherche particuliers.

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 ?

_____
_____
_____
_____
_____

- *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 ?

L'intervention d'un comité d'évaluation et de suivi des demandes d'accès serait uniquement justifiée si l'accès à des données de plateformes en ligne à des fins de recherche devenait réglementé et contraignant.

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** ?

Si l'accès à des données de plateformes en ligne à des fins de recherche devenait réglementé et contraignant, la présence desdites plateformes dans les comités d'évaluation et de suivi des demandes d'accès semblerait justifiée afin que ces plateformes bénéficient d'un droit de retour (et également d'une compensation financière pour la préparation et la mise à disposition du jeu de données couvrant ses frais de stockage et de transmission), ainsi que d'un droit de refus opposable pour protéger leur secrets commerciaux.

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 ?

_____
_____
_____
_____
_____

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 ?

_____
_____
_____
_____
_____

- *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 *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 ?

_____
_____
_____
_____
_____

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** ?

_____
_____
_____

[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.

_____
_____

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'Insee[15] ?

_____
_____
_____
_____
_____

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 ?

_____
_____
_____
_____
_____

- *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** ?

Nous ne sommes pas favorables à l'accès aux données des plateformes à des fins de recherche si les jeux de données contiennent des secrets industriels ou commerciaux ainsi que des données qui ne sont pas suffisamment dé-identifiées.

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** ?

Voir les réponses ci-dessus.

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 ?

Pour Criteo, le rôle des autorités de protection des données devrait se limiter au champ des données personnelles et à la protection de la vie privée.

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 ?

_____
_____
_____
_____
_____

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 ?

_____
_____
_____
_____
_____

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 ?

_____
_____
_____
_____
_____

- *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 ?

_____
_____
_____
_____
_____

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 ?

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

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

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