Réponse du groupe de chercheurs de l'Institut Mines-Télécom à la consultation ARCOM 2022 sur l'ouverture des données des plateformes dans le cadre du DSA

Chercheurs signataires:

Christine Balagué (IMT-BS, réseau Good in Tech), Inna Lyubareva (IMT Atlantique), Cécile Bothorel (IMT Atlantique), Nicolas Duminy (IMT Atlantique), Nicolas Soulié (IMT-BS), Annie Blandin (IMT Atlantique), Patrick Maillé (IMT Atlantique), Nicolas Jullien (IMT Atlantique)

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

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

Réponse :

Nous menons régulièrement des recherches utilisant des données issues d'une ou plusieurs plateformes en ligne, en raison de nos thématiques de recherche sur les biais des algorithmes des plateformes, le pluralisme de l'information en ligne, les stratégies d'influence via la publicité sur les plateformes durant les campagnes électorales, la désinformation, les dynamiques communautaires et la régulation des plateformes.

Plusieurs modes de collectes de données ont été utilisés pour ces recherches utilisant des données des plateformes, les principaux sont cités ci-dessous :

- Collecte par API pour Twitter, Instagram, Discord, Youtube, Europresse, Snapchat pour les différents projets de recherche, par exemple :

  ANR Pluralisme de l'Information en Ligne (PIL : 2018-2023)

  Thèse « Fouille de données temporelles et hétérogènes pour la modélisation et la détection des informations fallacieuses », co-direction Magic Lemp – IMT Atlantique R.Billot/I.Lyubareva

  Ecole Doctorale SPIN.

- Collecte manuelle pour Facebook (via API), notamment pour analyser les espaces d'interaction spécifiques destinés à l'aide à domicile et les médecins généralistes (Projet COAGUL : https://hal.archives-ouvertes.fr/hal-02915024/document).

- Collecte des données ouvertes des plateformes via API lors de la campagne présidentielle française (en utilisant des tokens d'accès + compte Facebook

developer) : https://www.goodintech.org/barometre-publicite-politique-campagne-electorale-2022.html

- Scraping de données sur Twitter, Instagram, Google Play, Transfermarkt, etc.
- Données Ulule : convention avec la plateforme qui nous a envoyé ses données. Procédé exemplaire pour les chercheurs qui a donné lieu à différentes publications, par exemple :

  Inna Lyubareva, Laurent Brisson, Cécile Bothorel, Romain Billot. Une plateforme de crowdfunding et son réseau social : L'exemple Ulule. Revue Française de Gestion, Lavoisier, 2020, Les mutations de l'accompagnement entrepreneurial, 46 (286), pp.135--151.https://doi.org/10.3166/rfg.2019.00402)

  Bothorel Cécile, Brisson Laurent, Lyubareva Inna (2021). How to Choose Community Detection Methods in Complex Networks to Study Cooperation and Successful Organizations. Computational Social Sciences. pp. 16-36. https://library.oapen.org/bitstream/id/b8dae25f-504a-4013-8a58-a11a62bc6f55/9788835124603.pdf

- Données d'audience Alexa, mobilisées dans différents travaux sur les modèles économiques numériques, par exemple :

  Inna Lyubareva, Fabrice Rochelandet, Yannis Haralambous (2020). Qualité et différenciation des biens informationnels. Une étude exploratoire sur l'information d'actualité. Revue d'économie industrielle, n° 172

- Données des moteurs de recherche : scraping des réponses d'une quinzaine de moteurs de recherche afin de les comparer (via des tests statistiques) et de détecter d'éventuels biais. Production d'un outil accessible au public sur www.snide.irisa.fr

- Les données géolocalisées, qui représentent un intérêt particulier pour la recherche et dont les plateformes sont de grands producteurs (cet aspect a été récemment discuté lors du conseil national de l'information géolocalisée).

- Développement d'une plateforme de collecte de données des plateformes de réseaux sociaux :
  Li Yingmin, Balagué Christine (2015), "Measure Social Metrics with Sodatech: a Monitoring and Analysis Platform of Big Data", ICWSM, May, Oxford

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

Réponse :

You Tube :

- Les quotas de base sont assez limités (10.000 requêtes/jour) et clairement insuffisants pour récupérer des bases de données de taille significative pour la recherche (à moins de passer des mois uniquement à récupérer les données). Le formulaire pour demander une augmentation de quotas est assez lourd et son processus de traitement est, jusque-là, assez obscur. Il serait bien de permettre un processus plus transparent pour les chercheurs. L'API You Tube n'est pas mal faite, mais il faudrait assurer sa stabilité (éviter de devoir changer le code des requêtes sur You Tube tout le temps). Le côté récursif des requêtes de l'API et l'échelle du volume de données à récupérer causent également des difficultés lors de la mise en place d'un processus de data mining robuste sur You Tube, car pour récupérer les commentaires il faut un id de vidéo, pour l'id de vidéo il faut une playlist, etc.
- Besoin de créer plusieurs comptes, de posséder plusieurs cartes Sim et de gérer les limitations de l'API, en particulier la reprise de la collecte après erreur. Difficulté de mettre à jour, i.e. rajouter des données après une première collecte. Pas d'accès aux données supprimées, ou aux changements d'alias.

Twitter:

- L'accès académique de Twitter fonctionne bien avec l'accès aux archives (sans limite dans le temps), recherche par mots clés/événements d'intérêt. Néanmoins, les délais de réponse de Twitter varient d'un cas à l'autre.
- Disparition de la plateforme Ads transparency de Twitter (accès pendant les campagnes électorales). Twitter a supprimé sa plateforme de transparence et a choisi de ne plus publier de publicités politiques. On peut néanmoins récupérer les données de 2018 à fin 2019, mais cela n'a plus d'intérêt aujourd'hui. Twitter ne fournit donc aucun moyen d'accéder au contenu des différentes publicités qu'ils proposent, alors que rien ne nous permet de vérifier que Twitter ne publie pas par inadvertance des publicités politiques.

Facebook :

- Difficulté à récupérer les posts et les fils de discussions, donc impossibilité d'analyser en profondeur, ou de façon statistique, les échanges, même quand les chercheurs sont membres d'un groupe Facebook (il faut être propriétaire du groupe pour réellement pouvoir collecter ces données)

· Obligation d'obtenir l'accord du créateur du groupe pour accéder aux données d'un groupe sur Facebook (*Benamar Lamya, Balagué Christine, Ghassany Mohamad (2017), The Identification and Influence of Social Roles in a Social Media Product Community. Journal of Computer-Mediated Communication, Volume22, Issue 6, pp. 337-362 http://onlinelibrary.wiley.com/doi/10.1111/jcc4.12195/full*)

Instagram :

· L' API d'Instagram ne permet pas d'avoir beaucoup de données, donc aujourd'hui on utilise plutôt du scraping et des faux profils

· Difficulté d'accès aux données sur les influenceurs

Moteurs de recherche:

· L'outil original a été développé pour monitorer plusieurs moteurs de recherche et récupérer les résultats, mais :

- on ne dispose quasiment pas d'API et donc il faut que notre outil parcoure la page de résultats et essaie d'en extraire les liens

- les moteurs de recherche changent régulièrement la mise en page et/ou la façon dont la page de résultat est codée, et donc on doit régulièrement adapter notre outil

Google : données publicitaires ouvertes lors des campagnes électorales

· Difficulté à prendre en main les données même pour faire un simple tri pour commencer. Les dates sont claires, mais pour ce qui est du pays de diffusion, les données sont loin d'être transparentes. Ainsi, des publicités de Google sur un candidat texan aux élections des gouverneurs aux USA auraient été prétendument publiées en France. Plus précisément, le fonctionnement réel de ciblage utilisé par les annonceurs semble difficilement compatible avec un filtrage effectif par pays : en effet, une publicité publiée sur un téléphone appartenant à un Américain en voyage en France compte-t-elle comme une publicité française ou américaine ? Si le public recherché est atteint, en termes de lecture sur le lieu, cela semble inapproprié, d'où le décalage lors du tri des données que l'on peut remarquer.

Snapchat :

· Très peu de données récupérables pendant les campagnes électorales (alors que données de publicité ouvertes)

Europresse :

· Absence de possibilité de faire une convention pour accéder aux API car n'est pas prévu par la plateforme

- Collecte de données actuelles non satisfaisantes

Remarques générales à toutes les plateformes :

- Asymétrie d'informations entre la plateforme et les chercheurs. En effet, on ne connaît pas toutes les données que possèdent les plateformes, et certaines données ne sont pas accessibles, comme par exemple certains commentaires supprimés sur des vidéos You Tube, ou certains posts supprimés sur Facebook ou Instagram.
- Faible quantité d'informations accessibles via les API, un nombre de requêtes possibles relativement faible, donc recours au scraping avec la création de faux profils
- Besoin de contrôler le déroulement du scraping: pour constituer des bases de données significatives, la collecte de données via des programmes automatisés implique de réaliser de très nombreuses requêtes auprès des sites. Ceci entraîne des blocages réguliers des faux profils, et la nécessité d'avoir un "pool" de faux profils. Les pages des sites web évoluent régulièrement, donc besoin de mettre à jour le code, etc. Un accès aux API avec la possibilité d'avoir plus d'autorisations (i.e. accès à plus de données, plus de requête, etc.) suite à une demande pourrait limiter ce type de problème
- Données d'audience ou même plus généralement : peu de détail sur le calcul précis et la définition de la donnée fournie. Ce manque d'explication sur la donnée rendue accessible rend plus complexe leur utilisation dans les travaux de recherche. On a besoin d'accéder à un volume important de données mais aussi à des explications détaillées et explicites sur leur signification (sinon les données n'ont aucune valeur).
- Données de publicités pendant les campagnes électorales : rien ne permettant d'analyser les publicités classées en-dehors de la catégorie politique, personne ne peut évaluer si des publicités politiques se retrouvent à tort dans de mauvaises catégories, ou si des publicités pour divers produits ne cacheraient pas un message politique. Dans ce cas, les chercheurs ne sont pas sûrs d'avoir les bonnes données pour analyser les enjeux publicité et désinformation pendant les campagnes électorales.

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.

Réponse :

- KissKissBankBank : recherche abandonner car pas d'accès aux API (en 2017) et les données fournies par la plateforme n'étaient pas exploitables. Ceci pose un problème plus général de préparation des données par les plateformes pour les

fournir aux chercheurs, nécessitant un travail conséquent de la part des plateformes, idem pour les API.

- Twitter affirme donner accès aux données pour les chercheurs, mais les réponses sont hétérogènes selon les chercheurs (un chercheur a reçu une réponse positive le jour même, tandis qu'un autre n'a toujours pas d'accès à Twitter 6 mois après sa demande).
- Facebook ne permet pas d'étudier des échanges; cela pour "protéger" les utilisateurs, ce qui est peut-être vrai, mais qui ne permet pas d'analyser les dynamiques collectives utilisant ce réseau
- Difficulté de faire des « testing » sur le marché du travail en utilisant les annonces publiées en ligne. Quel que soit la plateforme (Pôle Emploi, Monster, Indeed, etc.), les informations de contacts ne sont plus accessibles directement. Il faut postuler par l'intermédiaire de la plateforme sans connaître exactement la manière dont elle réalise la transmission des candidatures (filtrage, envoi direct, etc.).
- Les données de publicités lors des campagnes électorales sont de facilité d'accès très différente : les données Facebook permettent d'accéder de manière simple et de visualiser la publicité elle-même (reflet exact de ce qui a été publié), ce qui n'est pas le cas de Google. Facebook est donc priorisé dans l'analyse.

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.

Réponse :

- L'accès aux données de la plateforme Ulule est un cas très positif : données Ulule obtenues grâce à la collaboration avec la plateforme; convention bipartite ; plusieurs échanges par email avec des réponses à des demandes de précisions/vérifications sur les données ; accompagnement par la plateforme/ satisfaction d'Ulule par rapport à ces interactions.
- You Tube : malgré des quotas limités rallongeant plus que nécessaire la collecte de données, la documentation de l'API You Tube est très claire, et cette API est disponible dans de nombreux langages, chacun avec leur code d'exemple

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

Réponse :

Pas de politique de partage sur les données des plateformes à date.

Cependant, l'institut Mines-Telecom (IMT), pour les projets de recherche de ses chercheurs, a développé Terralab, plateforme « tiers de confiance » proposant des outils à l'état de l'art pour la collaboration entre entreprises et chercheurs destinée à l'accélération de projets IA, Big Data et IoT (https://www.imt.fr/recherche-novation/recherche/nos-partenariats/teralab/).

B. Gouvernance

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

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

Réponse :

- L'un des critères possibles d'accès selon les acteurs (et les projets) pourrait être le statut de l'employeur de l'acteur (but lucratif/non lucratif). Par exemple, un chercheur travaillant dans une institution publique à but non lucratif serait un cas différent d'un journaliste qui travaille pour son journal dont l'activité est à but lucratif.
- L'accès à des données utiles à la société pose plus globalement 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 via le DSA, la contribution des think tanks, des journalistes et de la société civile à la connaissance des problématiques liées aux plateformes en ligne mérite réflexion. La question de la neutralité des acteurs, au vu des financements qu'ils peuvent notamment recevoir de certaines plateformes, se pose également.

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

Réponse :

- Il semble que la réponse à cette question figure à l'article 40 du DSA
- Il pourrait être intéressant d'avoir une approche (plus) cohérente, dans les différents actes juridiques, du statut des chercheurs au regard de l'accès aux

données. Par exemple, le DSA prévoit un droit d'accès pour les chercheurs (sous certaines conditions), la directive copyright de 2019 a créé une exception au droit d'auteur pour la fouille de textes et de données au profit des chercheurs. Il y a d'ores et déjà des ponts entre les textes. Le DSA (art. 40) fait référence aux organismes de recherche tels que définis dans la directive de 2019. Pour être agréés, les chercheurs doivent être affiliés à ces organismes, que la directive définit comme suit : un «organisme de recherche», une université, y compris ses bibliothèques, un institut de recherche ou toute autre entité, ayant pour objectif premier de mener des recherches scientifiques, ou d'exercer des activités éducatives comprenant également des travaux de recherche scientifique:

a) à titre non lucratif ou en réinvestissant tous les bénéfices dans ses recherches scientifiques;

ou b) dans le cadre d'une mission d'intérêt public reconnue par un État membre; de telle manière qu'il ne soit pas possible pour une entreprise exerçant une influence déterminante sur cet organisme de bénéficier d'un accès privilégié aux résultats produits par ces recherches scientifiques;

- Il est évident aussi que les utilisateurs de ces plateformes ne veulent pas être "observés" et "jugés" par des chercheurs; par exemple, dans le cas du projet Coagul, les aides à domiciles, et notamment les personnes qui animaient les groupes, ne voyaient pas les chercheurs comme des acteurs neutres, mais comme des personnes d'un statut social (et culturel) plus élevé qui se posaient en observateurs/juges de leurs échanges.
- Au-delà des méta-données (qui échange avec qui) l'intérêt pour le chercheur est de plus en plus dans l'analyse fine des éléments de discussion (textes et espaces de commentaires) et cela pose des questions de confidentialité, qui semblent difficilement compatible avec la mise à disposition de jeux de données, autres que celles déjà publiquement disponibles/accessibles

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* ?

Réponse :

- Une solution serait de distinguer trois types de données :

  1/ Données d'intérêt général : accès public à tous sans restriction

  2/ Bases de données ouvertes à tous les chercheurs, utilisables aussi dans le cadre des enseignements dans le secondaire ou cours auprès d'étudiants

3/ Données plus sensibles et plus fine sur des projets plus spécifiques (accès limité sur demande d'une équipe de recherche)

- Il serait pertinent d'intégrer un niveau d'accès spécifique pour l'«enseignement », donc d'accorder l'accès aux données aux étudiants, en particulier en datascience, sciences sociales computationnelles, etc, dans le cadre de leur formation. Les modalités d'accès sont à préciser (passer par les enseignants-chercheurs – coordinateurs des formations ?)

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

i) Si oui :

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

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

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

ii) Si non :

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

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

Réponse :

- Si la solution retenue est un tiers de confiance, celui-ci devrait être l'autorité de régulation nationale.
- Nécessité de développer une plateforme qui centraliserait les demandes des chercheurs, les accès aux bases de données accessibles (cf réponse B2 : données d'intérêt général/ base de données ouvertes pour les enseignements)
- Pour les données qui existent en ligne (API) mais qui sont difficiles à collecter : il ne serait pas nécessaire de passer par un tiers de confiance, mais les plateformes devraient faciliter l'accès aux données : supprimer les quotas ; élargir la plage de temps accessible sur les données; garantir, dans la mesure du possible, la pérennité de l'API ; fournir les explications des données et indicateurs collectés; désigner un

interlocuteur côté plateforme ou animer une communauté d'aide en ligne sur les données collectées à l'image des forums open source.

- Nécessité d'éviter le goulot d'étranglement si mise en place d'un tiers de confiance. Plusieurs solutions sont envisageables :
  - Autorisation vis-à-vis des institutions académiques (certaines reconnues comme respectueuses des données personnelles, capables de stockage sur des serveurs sécurisés, capables d'anonymisation des données). Certaines institutions recevraient un « agrément » facilitant la procédure.
  - Au moment de la soumission du projet, fournir une liste de chercheurs de l'institution intéressés par ces données.
  - Guichet unique, pour des demandes de données sur plusieurs pays/ plusieurs plateformes (par exemple dans un projet européen).
  - Harmonisation du processus avec les autres textes sur l'accès des données pour les chercheurs (ex: exception de fouille de données et droit d'auteur; texte en préparation comme le *data governance act*)
  - S'inspirer des processus d'ouverture de données publiques (ex: CASD) et contribuer à l'ouverture de données d'intérêt général (qui seraient mises à disposition de tous)
- Si tiers de confiance, le process pourrait être le suivant :

  1/ Le chercheur remplit un formulaire en ligne sur une plateforme (description d'une page du projet plus une page de description des données).

  2/ Une fois la demande acceptée par le tiers de confiance, le chercheur accède aux données et il reçoit les coordonnées d'un contact au sein de la plateforme, pour répondre aux problèmes d'accès ou de compréhension des données. Pour l'accès aux données, le chercheur pourrait avoir un droit d'accès privilégié avec un Id spécifique pour faire sa requête sur l'API de la plateforme (cet ID lui permettrait de voir les variables disponibles, de cliquer sur les variables souhaitées, de paramétrer la requête puis de la lancer).
- Garder une procédure simple et agile
- Exiger une transparence de la sélection des chercheurs : la liste des projets et des raisons d'acceptation et de refus doivent être publiques et facilement accessibles à tous.
- Limiter le nombre de demandes par an par chercheur pour s'assurer que ce ne sont pas toujours les mêmes chercheurs qui ont accès aux données.
- Avoir une procédure spéciale (fast track) en cas de besoin d'agir sur des données en temps réel (ce qui est souvent le cas quand il s'agit des manipulations des informations/haine/etc dans le contexte politique/social instable)

· Engagement du tiers de confiance et des plateformes sur un délai court de réponse et d'accès aux données : si le délai entre la demande et l'attribution d'accès est trop important pour les chercheurs, comment concurrencer les travaux publiés par les plateformes elles-mêmes? Ce point complète la partie « Construction des projets scientifiques » dans lequel nous mentionnons l'asymétrie d'information entre les chercheurs et les plateformes, déjà existante, qui risque d'être renforcée.

· Possibilité de demander une mise à jour des jeux de données (recherches longitudinales)

· Accepter des demandes spécifiques sur des données « cachées » par les plateformes (ex: projets effacés/ accès aux users You Tube bannis)

· Créer une plateforme de partage entre chercheurs : partage d'expérience des chercheurs (très important) ; partage de code pour reproductibilité des expériences…

· Apporter des réponses aux questions suivantes : à qui appartiennent les données ? jusqu'à quand peut-on les conserver ? Est-ce qu'on peut les réutiliser dans un autre projet ou pour l'enseignement ? Quelles données peut-on publier (pour certaines revues, fournir les données et publier le code sont obligatoires) ? Sur les publications, est ce que les plateformes ont leur mot à dire ?

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

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

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

Réponse :

· L'autorité de régulation devrait s'assurer que les plateformes ont bien donné l'accès aux données pour le chercheur. Ceci peut s'effectuer sur la plateforme qui centralise les demandes, en affichant en temps réel l'état d'avancement, depuis le dépôt de la demande jusqu'à l'accès réel aux données (plusieurs étapes pourraient être paramétrées).

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

Réponse :

· Rendre publiques les demandes d'accès aux données des chercheurs et les réponses des plateformes, comme sont publiques les réponses des plateformes aux questionnaires sur la désinformation de l'ARCOM. · L'autorité de régulation en tant que tiers de confiance dit oui

ou non aux projets des chercheurs. Sur certaines demandes, les plateformes pourraient réagir en précisant que la demande ne rentre pas dans le cadre du DSA, ou que les données demandées

· Rendre publiques les recherches (avec accès aux données) financées par les plateformes elles-mêmes (il se peut qu'avec le DSA les plateformes financent elles-mêmes certains chercheurs pour travailler sur leurs données…).

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

Réponse :

· Les plateformes doivent avoir un rôle de partenaire et d'accompagnement des chercheurs mais non de censure.

poseraient des problèmes d'atteinte à la privacy, et ce serait alors à l'autorité de régulation de trancher.

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

Réponse :

· Complexité de la demande et du système
· Asymétrie de moyens entre les plateformes et les chercheurs
· Risque de favoriser les publications par les plateformes (qui ont des chercheurs en interne) et non par des chercheurs plus indépendants, et donc un brouillage de l'information diffusée (biais de recherche en faveur des plateformes, à l'image de ce qui peut exister dans les recherches médicales)

C. Construction des projets scientifiques

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

Réponse :

· Au-delà d'un certain seuil de sollicitation, la plateforme devrait organiser un séminaire en ligne (1h) de présentation des données, de leur transmission/collecte, etc.. Des vidéos en ligne accessibles aux chercheurs pourraient être élaborées par les plateformes.
· Favoriser l'interaction directe chercheurs-plateformes comme modèle complémentaire afin de, avec le temps, réduire l'asymétrie de l'information et

mieux définir les collaborations envisageables, les difficultés et pistes d'évolution entre le monde académique et les plateformes, dans une perspective bottom-up.

- Réfléchir à des différentes modalités d'interactions plateformes-chercheurs (exemple : journée d'étude spécialisée, séjour des chercheurs au sein des plateformes, séminaire de discussion plateformes/chercheurs sur l'accès technique aux données, ou sur la compréhension des données)

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.

Réponse :

- Difficile de limiter les thèmes des projets de recherche. Les plateformes impactent de nombreux domaines qu'ils soient relatifs à la vie privée ou professionnelle des personnes avec des conséquences économiques et sociétales très variées.

- Sans limiter les thèmes, un critère d'évaluation des demandes pourrait porter sur leur apports et enjeux pour les missions d'intérêt général identifiées comme prioritaires (sans que le projet porte spécifiquement sur les thèmes des missions).

- Le DSA mentionne un périmètre de recherches sur des risques systémiques. Ces risques systémiques peuvent évoluer dans le temps, le périmètre doit donc rester large. Par exemple, l'accession aux données publicitaires est nécessaire pour étudier la manipulation de l'opinion et la désinformation (les données publicitaires permettraient de mesurer et de comprendre l'amplification ou la modération de certains contenus).

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 ?

Réponse :

- Les réponses à cette question ont été données en B.3

- Prévoir un Fast track sur des thématiques imprévues où les données doivent être étudiées en temps réel (ex : Guerre en Ukraine)

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

Réponse :

· Il est indispensable de ne pas complexifier la procédure (s'il y a un tiers de confiance, c'est lui qui devrait décider ou non de la demande d'accès. Dans ce cas il n'est pas nécessaire d'avoir un comité d'évaluation et de suivi).

· Il serait pertinent d'avoir un suivi de l'état d'avancement de la demande en temps réel sur un site qui centraliserait les demandes.

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

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

Réponse :

· Comme déjà évoqué, nous suggérons la création d'une plateforme centralisant et traçant l'évolution des demandes, les publications, le suivi de cette ouverture progressive des données, les règles de fonctionnement (y compris de citation de la plateforme)

· Définir les recours pour les chercheurs, et informer ceux-ci de leur possibilité de recours

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** ?

Réponse :

· Selon nous, le droit de retour de la plateforme pourrait consister en une demande de modification de la demande d'accès (par exemple si la demande ne respecte pas le RGPD) ou de refus si la demande ne rentre pas dans le cadre du DSA. Dans ces cas, l'autorité de régulation devrait au final trancher.

· On peut regretter que l'encadrement juridique des cas pour lesquels il y aurait un refus n'est pas précisée dans le DSA.

· Il semble nécessaire de préciser les conditions de droit de refus des plateformes, et de rendre accessible publiquement toute demande et son refus.

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

Réponse :

Quelques suggestions :

- Équipes interdisciplinaires, dont les recherches s'inscrivent dans les thématiques d'intérêt général (ou contribuent indirectement à celles-ci) et s'engageant à diffuser non seulement les résultats sous forme de publication, mais également à mettre à disposition des outils open source (pour la collecte/stockage/analyse), des bases de données et les méthodologies développées afin de favoriser l'utilisation des données et des résultats par d'autres disciplines.
- Possibilité de collaboration avec d'autres équipes demandant l'accès aux mêmes données ou aux données complémentaires (peut être un critère intéressant)
- Ne pas limiter l'accès aux structures implantées dans au moins 2 pays de l'Union Européenne (rend tout de suite la démarche plus complexe)
- Pas de critère de taille des équipes de recherche (pour intégrer des petites nouvelles équipes)
- Capacité des équipes ou des institutions employant les chercheurs à garantir la confidentialité et la sécurité des données recueillies (infrastructure d'hébergement, de traitement, etc…).
- Il serait intéressant d'établir une liste de critères à "cocher" lors de la demande, comme premier filtre automatique, par les demandeurs d'accès aux données, pour qu'ensuite l'autorité de régulation vérifie en comité l'identité des chercheurs et la pertinence du projet.

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 ?

Réponse :

- La dimension temporelle ne doit pas être un critère excluant.
- En revanche, une fois l'accès aux données accordé, il serait préférable de limiter le moins possible cet accès dans le temps.
- Favoriser la durée longue d'accès aux données (car coût fixe de mise en place des infrastructures de téléchargements, de stockage etc…) donc prévoir une procédure allégée d'accès à l'actualisation des données (procédure de renouvellement très

allégée). Ceci permettrait de préserver la continuité et la pérennité des travaux, de profiter des rendements d'échelle (il y a toujours un coût fixe important d'installation et d'appropriation de l'infrastructure pour recevoir/stocker/organiser les données) et de pouvoir agir en temps réel.

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 ?

Réponse :

· Une telle certification alourdirait le processus.

· La rigueur scientifique est déjà évaluée par les comités scientifiques des conférences et des revues (généralement plusieurs rounds de reviewing sont nécessaires pour la publication dans un journal de premier rang).

· Pas de certification externe mais, par exemple, obligation de citer la plateforme-partenaire et le dispositif d'accès aux données dans la publication (comme obligation de citer l'ANR dans les publications financées par l'institution). Il y aura plus de visibilité pour ensuite pouvoir évaluer et comparer dans quelle mesure différentes plateformes « jouent le jeu » (avec le nombre de leur citation dans les publications/communications)

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** ?

Réponse :

· Les études basées sur la demande d'accès aux données sont généralement des études statistiques, dont les résultats sont des résultats agrégés, ce qui limite les problèmes.

· Cependant, la publication des résultats (obligation du DSA) peut poser dans certains cas des problèmes d'incompatibilité avec la sensibilité des données. Cette question est relativement classique en recherche, et pas uniquement liée aux plateformes (ex : données de santé, terrains sensibles etc…). Dans plusieurs institutions de recherche, ce type de recherche sur des données sensibles doit être validée par le comité d'éthique de l'institution académique du chercheur. On peut imaginer que les demandes d'accès aux données soient accompagnées de l'avis d'un comité d'éthique, interne pour les institutions en disposant, et à définir pour celles qui n'en ont pas. De même, la CNIL a un cadre réservé à la recherche pour l'usage des données numériques sensibles, qui pourrait être utilisé.

· Données sensibles pour les plateformes : définir les règles de ce qu'on peut publier dans la convention bilatérale entre chercheur et plateforme

· On pourrait aussi prévoir d'interdire de nouvelles demandes d'équipes qui n'auraient pas respecté les principes de confidentialité

D. Protection des données et considérations techniques

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

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

Réponse :

· Préciser sur la plateforme de paramétrage de la demande d'accès aux données 2 niveaux :

- Un niveau de demande « classique » pour laquelle le chercheur coche des types de données/ durée etc… par plateforme
- Un niveau plus spécifique (plutôt réponse ouverte) dans lequel le chercheur explicite une granularité supplémentaire ou ses demandes précises

· Envisager des solutions de traitement des données à distance lorsque les données sont très sensibles (les données ne sortiraient pas des plateformes dans ce cas, on un modèle de type CASD)

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 ?

Réponse (déjà fournie en B.3) :

· Créer une plateforme pour chercheurs avec partage : des données des plateformes ; de l'expérience des chercheurs (très important) ; code pour reproductibilité des expériences…

· Cette plateforme devrait être ouverte à tous les chercheurs qui veulent construire des indicateurs ou mesures innovants.

· Imposer une phrase spécifique dans les articles pour repérer tous les papiers publiés avec ouverture des données des plateformes, pour avoir de nouveaux indicateurs et des méta-analyses.

D.2. Les accès aux données peuvent-ils et doivent-ils faire l'objet d'une **co-construction** à part égale entre acteurs de la gouvernance, chercheurs et plateformes sur le modèle du CASD de l'Insee15 ?

Réponse :

- L'exemple du CASD est pertinent (déjà cité dans nos réponses)
- En plus du CASD, nous suggérons une plateforme de partage d'expériences entre chercheurs (cf description dans la réponse B.3)

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 ?

Réponse :

- Garder un cadre suffisamment large et souple pour qu'il puisse s'appliquer aux évolutions régulières des plateformes (à l'image du DSA).
- Solution de traiter les données à distance lorsque les données sont très sensibles.

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 ?

Réponse (déjà fournie en B.3) :

- Le chercheur devrait avoir un droit d'accès privilégié avec un Id spécifique pour avoir une requête sur l'API de la plateforme (voir les variables disponibles, cliquer sur les variables souhaitées, paramétrer la requête puis la lancer pour accéder aux données).
- Avoir ensuite un contact au sein de la plateforme et à la communauté des autres utilisateurs, qui partagent leur expérience, pour toute question de précision sur les données ou de problème technique d'accès.

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

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

Réponse :

- Héberger les données sur une plateforme sécurisée dédiée aux chercheurs (exemple : CASD ou encore plateforme Teralab développée à l'Institut Mines Telecom : https://www.imt.fr/recherche-innovation/recherche/nos-partenariats/teralab/).
- Fournir un agrément (comme pour les hébergeurs de données de santé) à des hébergeurs de données, lié au degré de sécurité du stockage des données.
- Définir différents niveaux de sensibilité et processus (de la collecte jusqu'au stockage).

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 ?

Réponse :

- En cas de projet utilisant des données sensibles, dépôt de projet auprès de la CNIL (procédure spécifique aux projets de recherche déjà existante).
- Éventuellement validation du projet par le comité d'éthique de l'institution académique du chercheur.

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 ?

Réponse :

- La question importante est selon nous plutôt le financement par la France et par l'Europe des projets de recherche utilisant les données des plateformes, dès la mise en place du DSA. Comment s'assurer que des financements importants de la recherche seront « fléchés » sur des projets utilisant des données des plateformes ?
- Avoir une information sur les financements potentiels de ce type de projets sur la plateforme centralisée de demande d'accès aux données des plateformes serait nécessaire.

E. Faisabilité de l'accès et incitations

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

Réponse :

· Le dépôt d'une demande devrait intégrer les éléments relatifs au RGPD en fonction des données demandées et de leur granularité.

· S'inspirer du processus de la CNIL pour un accès de données sensibles aux chercheurs

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 ?

Réponse :

· Si les risques sont systémiques, le niveau de financement global de projets utilisant des données des plateformes devrait être suffisamment important pour couvrir les besoins des différents chercheurs…

· La mise en place d'une plateforme de partage des expériences (y compris techniques) des chercheurs permettrait de réduire les écarts entre institutions académiques, et de faciliter la courbe d'apprentissage des plus petites équipes de recherche (point déjà évoqué en réponse B.3).

· La simplification d'accès aux données va contribuer à diminuer les écarts et à démocratiser l'utilisation des données (en recherche et en enseignement, si autorisé)

· Une interaction directe, via un interlocuteur, entre la plateforme et les chercheurs permettraient aux chercheurs avec moins de capacités techniques d'avoir des réponses aux questions qu'ils se posent et de monter en compétence.

· Développement progressif d'interaction directe, via différentes modalités (cf.C1), entre les plateformes et les chercheurs va également contribuer à cette tâche.

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 ?

Réponse :

· Reconnaissance envers les plateformes qui ouvrent leur données (visibilité des retours satisfaisants des chercheurs, visibilité des projets, visibilité des bons échanges chercheurs-plateformes)

- Ceci pourrait s'effectuer sur le site qui centraliserait les demandes mais aussi dans des communications officielles de l'autorité de régulation (un bilan annuel sur l'ouverture des données des plateformes, avec publication des demandes, des accès effectifs, etc…) ou des institutions académiques utilisant les données des plateformes.
- Reconnaissance envers les plateformes liées au nombre de publications générées par l'ouverture des données. Ceci pourrait donner lieu à une évaluation des différentes plateformes sur des critères comme le nombre de demandes acceptées, sur le nombre de publications, etc…

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** ?

Réponse :

- Ne pas multiplier les comités au risque d'avoir des durées très longues d'accès aux données.
- Simplifier au maximum la demande des chercheurs

Annexe 1 :

Quelques références de travaux et publications des chercheurs signataires de cette réponse à la consultation ARCOM :

Balagué Christine, Fayon David (2022): *Pro en réseaux sociaux*, Editions Vuibert, Paris, ISBN : 978-2-311-62468-7

Balagué Christine, Jules Rostand, Quentin Michaud (2021), Stratégies d'influence et publicité en ligne pendant la campagne présidentielle française, https://www.goodintech.org/barometre-publicite-politique-campagne-electorale-2022.html

Balagué Christine (2020), Explicabilité des algorithmes des plateformes digitales, *Audition par le CNCDH* (Commission Nationale Consultative des Droits de l'Homme), 18 déc

Balagué Christine (2019), Réseaux sociaux et communautés de patients, 21 février, audition par la *Haute Autorité de Santé, HAS*

Benamar Lamya, Balagué Christine, Ghassany Mohamad (2017), The Identification and Influence of Social Roles in a Social Media Product Community. *Journal of Computer-Mediated Communication,* Volume22, Issue 6, pp. 337-362 http://onlinelibrary.wiley.com/doi/10.1111/jcc4.12195/full

Balagué Christine , Renard Damien (2015), "Understanding consuming, contributing and creating behaviors on social networks", *EMAC, May, Leuven, Belgium*

Blandin Annie, "La question concurrentielle dans le contexte des états généraux des nouvelles régulations numériques", *Contrats concurrence consommation*, LexisNexis, 2019, Les droits de la concurrence d'une réforme à l'autre. ⟨hal-03337083⟩

Blandin Annie, "La protection de l'individu face à l'automatisation de la présentation des contenus par les plateformes", *Etudes digitales*, Classiques Garnier, 2019, Les plateformes, 2 (8), ⟨10.15122/isbn.978-2-406-10497-1.p.0027⟩. ⟨hal-03335902⟩

Blandin Annie, "Digital Services Act et Digital Markets Act : un nouveau cadre européen pour la régulation des plateformes", à paraître dans un ouvrage sur la politique européenne du numérique.

Bothorel Cécile, Brisson Laurent, Lyubareva Inna (à paraître en 2023). Plateformes en ligne et analyse des dynamiques communautaires. In « Diversité des approches méthodologiques en sciences sociales » (Dir. I.Lyubareva et R.Waldeck), ISTE

Bothorel Cécile, Brisson Laurent, Lyubareva Inna (2021). How to Choose Community Detection Methods in Complex Networks to Study Cooperation and Successful Organizations. *Computational Social Sciences*. pp. 16-36.

Boudjani Nadira, Haralambous Yannis, Lyubareva Inna (2020). Toxic Comment Classification For French Online Comments. IEEE International Conference on Machine Learning and Applications (IEEE ICMLA)

Li Yingmin, Balagué Christine (2015), "Measure Social Metrics with Sodatech: a Monitoring and Analysis Platform of Big Data", *ICWSM, May, Oxford*

Lyubareva Inna, Brisson Laurent, Bothorel Cécile, Mesangeau Julien, Boudjani Nadira, Haralambous Yannis (Article en préparation). Pluridisciplinary study of echo chambers in comments on Youtube: example of the French media channels.

Lyubareva Inna, Mesangeau Julien, Boudjani Nadira, El Badisy Imad,Brisson Laurent (2021). La plateformisation des médias français et le ton du débat public : exemple de Youtube. *Revue Communication (Collection Actes)*, 38 (2), ⟨10.4000/communication.14433⟩ (Open access) https://doi.org/10.4000/communication.14433

Lyubareva Inna, Brisson Laurent, Bothorel Cécile, Billot Romain (2020). Une plateforme de crowdfunding et son réseau social : L'exemple Ulule. *Revue Française de Gestion*, Lavoisier, Les mutations de l'accompagnement entrepreneurial, 46 (286), pp.135--151. https://doi.org/10.3166/rfg.2019.00402

Maillé, Patrick (2022). Un biais dans le moteur, *Pour la Science*, no 541, https://www.pourlascience.fr/sr/article-partenaire/un-biais-dans-le-moteur-24363.php

Manant M., Pajak S. and Soulié N., 2019, "Can social media lead to labor market discrimination? Evidence from a field experiment", *Journal of Economics and Management Strategy*, vol. n°28 (2), https://doi.org/10.1111/jems.12291

Matthews Jean-Marie, Cardon Dominique, Balagué Christine (2022**).** From Reality to World. A Critical Perspective on AI Fairness. *Journal of Business Ethics,* DOI : 10.1007/s10551-022-05055-8