Arcom

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

META

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

INTRODUCTION

Nous apprécions l'occasion de répondre à cette consultation sur l'accès aux données des plateformes en ligne pour la recherche. Nos réponses s'appuient sur l'expérience et l'expertise de Meta en matière de partage de données à des fins de recherche et de transparence. Elles s'appuient en particulier sur notre travail de partage de données avec des chercheurs académiques depuis deux des plus grandes plateformes au monde, Facebook et Instagram, et sur notre participation au Groupe de travail de l'Observatoire européen des médias numériques (*European Digital Media Observatory*) (EDMO) sur l'accès des chercheurs aux données des plateformes.

Cette expérience a permis de tirer trois grands enseignements qui, selon nous, peuvent rendre les futurs exercices de partage de données plus fructueux.

1. **Des organes de gouvernance externes indépendants sont nécessaires.** Un tel organe est essentiel au fonctionnement efficace et équitable des nombreux aspects d'un écosystème de partage de données, y compris les questions fondamentales de gouvernance (qui obtient quoi, quand et comment), le fonctionnement équitable de cette gouvernance lorsqu'il y a de multiples parties intéressées, l'interprétation des obligations juridiques, la hiérarchisation des besoins académiques/du public et l'élaboration de nouvelles mesures de protection de confidentialité et de sécurité, y compris des conseils sur le moment opportun de déployer de telles mesures de protection.
2. **Les « produits » de données polyvalents sont plus durables et offrent des possibilités de recherche à plus long terme.** La mise à disposition de données pour les chercheurs exige des efforts de la part des équipes d'ingénierie, affaires publiques, service juridique et partenariats, ce qui signifie qu'il y aura nécessairement des coûts d'opportunité pour la publication des jeux de données, nécessitant une hiérarchisation par les parties intéressées pour s'assurer que ce qui est finalement conçu soit sert le grand public au sens le plus large soit soutient la recherche ayant un impact maximal. Se concentrer sur des jeux de données polyvalents qui répondent aux priorités identifiées par l'organe indépendant susmentionné procure un avantage important à la communauté scientifique dans son ensemble et à l'intérêt public, plutôt qu'à un petit groupe de chercheurs influents ou d'élite.
3. **La recherche sur les fausses informations ou la désinformation est importante, mais la production de jeux de données sur le sujet est de fait privilégiée par rapport à toutes les autres possibilités relatives à l'impact sur l'intérêt public.** En collaboration avec la communauté académique, il pourrait être possible d'étudier une

multitude de questions de sciences sociales à l'aide de données provenant de plateformes technologiques qui éclairent profondément les politiques publiques fondées sur les données. Par exemple, nos données sur les Zones de trajets domicile-travail peuvent aider à développer une meilleure planification urbaine. En outre, le développement des jeux de données n'est pas le seul moyen d'informer le public à propos de l'activité sur les plateformes. En élaborant en collaboration des normes de rapport et d'audit, les entreprises peuvent publier des rapports que les chercheurs peuvent effectivement utiliser, valider et auxquels ils peuvent faire confiance.

Nous avons répondu plus en détail aux questions spécifiques ci-dessous et apprécions l'occasion de partager notre expérience. Nous serions heureux d'avoir des échanges supplémentaires sur ces sujets et vous invitons à nous contacter si vous avez d'autres questions.

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

Questions spécifiques à destination des plateformes en ligne :
A.5 Avez-vous établi une politique de partage de vos données avec des tiers à des fins de recherche ?

i) Si oui :
- *depuis quand existe-t-elle ?*
- *concerne-t-elle une ou plusieurs catégories de bénéficiaires particuliers (chercheurs, ONGs, entreprises, etc.) ?*
- *existe-il des critères de sélection de ces bénéficiaires ? Si oui, lesquels ?*
- *quel(s) type(s) de données cette politique concerne-t-elle ?*
- *intègre-t-elle un volet de contrôle ou de suivi de l'utilisation qui est faite des données délivrées ?*

ii) Si non, quelles sont les raisons pour lesquelles vous n'avez pas initié une telle politique ? Il peut notamment s'agir de risques d'ordre juridique, réglementaire, technique, financier, etc Précisez quelle a été votre évaluation de ces risques menant à la décision de ne pas ouvrir vos données.

Nous avons activement participé au Groupe de travail de l'EDMO sur le partage des données des plateformes aux chercheurs afin de développer des processus standardisés pour partager les données avec les chercheurs. Nous pensons qu'un code de conduite clair à suivre par les plateformes et les chercheurs est essentiel pour équilibrer efficacement une volonté de transparence et de recherche avec la protection des données à caractère personnel. En l'absence d'une réglementation claire définissant les règles pour la recherche et le partage de données de manière transparente, et imposant des structures de responsabilité qui suivent les données partagées, nous recommandons une approche fondée sur le risque pour le partage des données avec les chercheurs.

Nous disposons d'un cadre interne pour le partage des données avec les chercheurs, qui est partiellement déployé avec des équipes, et travaillons à son amélioration et à son déploiement à plus grande échelle. Nous avons également mis en place des politiques et des procédures pour régir toutes les pratiques de partage des données conformément au RGPD et au FTC Consent Order.

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

Bien que nous comprenions le désir d'élargir la définition des chercheurs au-delà de ceux qui sont rattachés aux universités, nous devons d'abord résoudre deux questions liées à un accès plus large : des critères d'inclusion clairs et une responsabilité institutionnelle.

- Les critères d'inclusion constituent un défi opérationnel : les entreprises ne sont pas (et ne devraient pas être) en mesure de définir quels membres de la société civile sont qualifiés. Les organisations de la société civile peuvent prendre une myriade de configurations et de formes institutionnelles – en effet de nombreux acteurs importants ne sont pas des organisations du tout mais plutôt des individus. Identifier ceux qui sont compétents pour soutenir la recherche scientifique n'est pas une expertise présente dans la plupart des entreprises, voire dans aucune. En outre, dans le cas des plateformes de réseaux sociaux, il existe une tension nécessaire et utile entre celles-ci et certaines composantes de la société civile. Beaucoup de ces groupes existent pour plaider en faveur du changement, et eux-mêmes sont souvent d'avis opposé, ce qui

signifie qu'il y aura toujours une tension nécessaire (et utile !) entre des composantes de la société civile et les plateformes de réseaux sociaux. Les universités sont, en revanche, beaucoup plus simples à identifier et sont presque toujours accréditées par d'autres organismes, notamment les gouvernements eux-mêmes. En outre, l'un des principes fondamentaux de la recherche scientifique moderne est l'importance de revue par les pairs, que presque toutes les universités connaissent bien et pour lequel elles disposent de processus. Tel n'est pas le cas dans les établissements autres que ceux de recherche.

- Plus complexe, cependant, est la question de la responsabilité institutionnelle. Il y a de nombreux avantages, tant pour les chercheurs que pour les plateformes, à travailler avec les chercheurs des universités.
    - Les universités ont leurs propres codes de conduite éthiques et comités de revue et offrent une protection supplémentaire aux données et à leur utilisation. Il s'agit en fait de garanties organisationnelles supplémentaires qui peuvent intervenir conjointement avec les mécanismes de surveillance et d'application que les plateformes peuvent mettre en place, et/ou, aux fins de l'indépendance de la recherche, peuvent remplacer les mécanismes des plateformes, le cas échéant.
    - Les institutions ont plus de ressources et de pouvoir de négociation que les chercheurs individuels.
    - Les universités sont souvent des entités juridiques constituées de longue date qui peuvent assumer la responsabilité de l'activité du chercheur, à la fois en protégeant les chercheurs individuels et en fournissant aux entreprises et aux régulateurs une forme institutionnelle à laquelle ils peuvent demander réparation.

De nombreuses initiatives, dont notamment le Code européen de bonnes pratiques contre la désinformation, encouragent la mise en place de mesures pour inclure des représentants d'organisations à but non lucratif et de la société civile. Cependant, cette approche met en évidence à la fois des questions juridiques/éthiques, en plus des questions opérationnelles. Notamment lorsque la définition des chercheurs est plus large que celle des chercheurs affiliés à une université, il est essentiel de mettre en place un processus de sélection des chercheurs qui garantisse qu'ils mènent un projet approuvé et respectent les exigences en matière de protection des données.

- Aux États-Unis, la meilleure façon de surveiller cette situation est de recourir à des IRB (Comités de revue institutionnels) affiliés à une université ou à un organisme semblable à la NSF (Fondation nationale pour la science) pour l'examen des projets de recherche.
- En Europe, le Groupe de travail de l'EDMO relatif à l'Article 40 travaille à un processus d'examen et d'approbation des projets de recherche qui utiliseraient un organe *ex ante* remplissant une fonction similaire à la NSF.

La procédure de contrôle doit garantir que les chercheurs sont formés et capables à la fois d'exécuter leurs projets de recherche proposés et de protéger toutes les données nécessaires. De plus, cette procédure d'examen peut aider à réduire la probabilité que le partage des

données de recherche soit utilisé comme porte dérobée à d'autres finalités non admissibles. Demander à tout chercheur de s'affilier à une université pour être approuvé pour un projet de recherche pourrait être un moyen efficace d'assurer un contrôle approprié. Si un autre acteur ne réussit pas le processus d'examen pour les « chercheurs », il peut être plus approprié pour eux de recevoir uniquement des données moins à risque en termes de données à caractère personnel (c'est-à-dire des informations manifestement publiques) et/ou dans des formats plus protecteurs des données à caractère personnel (c'est-à-dire des rapports et des données agrégées plutôt que des données brutes).

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 ?

Dans certains contextes limités et spécifiques ou sujets ayant un intérêt public important, les plateformes peuvent être tenues de rendre public des données d'une manière à protéger les données à caractère personnel (par exemple, des publicités politiques menant à des élections clés). En outre, les rapports sur les chiffres et la transparence peuvent être utilisés pour assurer une plus grande transparence sans augmenter les risques pour les données à caractère personnel.

Les « données brutes » ne sont pas le seul moyen de permettre la recherche ou de créer suffisamment de transparence pour permettre au public d'enquêter. Des rapports normalisés sur la transparence, assortis de mécanismes d'audit appropriés, pourraient s'étendre à l'ensemble de l'industrie et être réalisés de manière à favoriser la recherche, d'une manière similaire à celle utilisée pour la publication de demandes d'accès gouvernementales ou la suppression de contenus. Cela permettrait aux universitaires, aux journalistes et aux défenseurs de causes d'effectuer des recherches comparatives entre les plateformes.

La réglementation pourrait également établir des groupes de travail pour le secteur qui détermineraient les normes appropriées que les entreprises peuvent respecter lorsqu'elles publient volontairement des indicateurs. Pour ce faire, les plateformes pourraient notamment indiquer combien de chercheurs ont accès aux données par le biais de leurs outils et le volume de données mises à disposition. Les plateformes pourraient également compter le nombre de publications qui utilisent leurs jeux de données.

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 ?

Nous pensons qu'un modèle hybride est préférable, dans lequel un organe tiers indépendant prend des décisions de gouvernance (qui obtient quoi, quand et comment) sans traiter les données directement. Cette perspective est directement basée sur notre travail avec le Groupe de travail de l'EDMO sur l'accès aux données et les exigences du RGPD. Si un tiers traitait lui-même des données, il serait responsable de la protection des données de ces utilisateurs, y compris dans certains cas en tant que responsable du traitement des données.

Cela dit, nous pensons également qu'un rôle clé de l'organe de décision en matière de gouvernance est d'établir des normes claires pour les mesures de protection techniques et, une fois définies, il est facile d'imaginer que de multiples entités pourraient distribuer des données, y compris les universités elles-mêmes, tant qu'elles sont liées à ces restrictions.

En ce qui concerne la nature de ces garanties et la manière dont elles devraient être mises en œuvre, nous renvoyons une fois de plus au travail accompli par le Groupe de travail de l'EDMO sur l'accès aux données. Son dernier rapport explique comment évaluer conjointement les risques liés à l'échange de données et aux activités de recherche, puis comment appliquer des garanties qui correspondent au niveau de risque évalué.

B.4 Dans l'hypothèse d'un mode de régulation qui impliquerait l'intervention d'un tiers de confiance dans l'ouverture des données pour des projets de recherche :
i) qui aurait la charge de contrôler la mise en œuvre du protocole de demande ?

ii) quels garde-fous pourraient être mis en place pour assurer un accès à des données permettant de répondre au besoin exprimé de manière satisfaisante ?
iii) comment la transparence des décisions des organisateurs du protocole d'accès devrait-elle être garantie ?
iv) quelle place et quels rôles devraient avoir chacune des parties prenantes et notamment les plateformes ?
v) identifiez-vous des risques inhérents à ce modèle ? Lesquels ?

En plus de notre explication des avantages des intermédiaires à la question B3, il est également fait référence au cadre d'évaluation des risques développé par le Groupe de travail de l'EDMO

sur l'accès aux données (voir en annexe de ce document). Ce cadre évalue le risque associé à la fois aux données à partager et à l'activité de recherche elle-même afin de s'assurer que les mesures appropriées sont maintenues.

En ce qui concerne l'application, nous maintenons également qu'un intermédiaire indépendant est en bonne position pour assurer le respect des mesures. Un intermédiaire indépendant en matière de gouvernance est également dans une position unique pour soutenir les efforts de transparence : s'il existe un organe unique à l'échelle de l'UE qui approuve les projets, il est en mesure de publier des informations sur les recherches en cours et passées sur son site internet. La centralisation de ces informations rend la transparence plus efficace, car le public peut y accéder facilement.

Bien que nous pensons que les plateformes ont un rôle important à jouer dans un tel arrangement, nous pensons que c'est davantage dans la gouvernance d'un tel organe que dans le fonctionnement d'un tel organe. Du point de vue de la gouvernance, nous pensons qu'il y a un avantage important à ce que l'expérience et l'expertise des plateformes jouent un rôle. Par exemple, un tel représentant comprendra mieux quelles sont les demandes de données qui sont probablement plus complexes que les autres, et devrait également être en mesure d'aider les autres à comprendre quelles données de l'activité sont plus propices à la recherche que d'autres.

C) Construction des projets scientifiques

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

Les plateformes pourraient être tenues de publier des livres de code expliquant les points de données disponibles, similaires à ceux exigés par le Code européen de bonnes pratiques contre la désinformation et par le projet de Code de Conduite de l'EDMO en application de l'Article 40 du RGPD. Le libellé ci-dessous est extrait du projet de Code de Conduite de l'EDMO :

« 2.1 Contenu des Livres de codes
Un Livre de codes comprendra, au minimum :

a. Une description des catégories de données contenues dans l'ensemble de données (les « Champs de données »), y compris la présence de toute donnée qui concerne les Articles 9 ou 10 du RGPD ;

b. Une description des catégories et du nombre approximatif de personnes concernées représentées dans l'ensemble de données ;

c. Une description de ce que représente l'ensemble de données et de son adéquation à la recherche, y compris : (i) l'exhaustivité de l'ensemble de données (par exemple, par rapport aux données disponibles détenues par le Responsable de la sécurité des données (*DSO*) ou à une autre définition de la population concernée), (ii) l'exactitude de

l'ensemble de données (par exemple, la mesure dans laquelle les données qu'il contient sont connues pour être vraies, à la fois en spécificité et en précision) ; et (iii) le caractère actuel de l'ensemble de données (par exemple, la mesure dans laquelle les données représentent la réalité à un moment donné) ;

d. Une liste du ou des pays dans lesquels se trouvent les personnes concernées représentées dans l'ensemble de données, s'ils sont connus ;

e. Une description de tous les paramètres pertinents de confidentialité ou autres qui s'appliquent aux données, y compris, sans s'y limiter, les paramètres sélectionnés par les personnes concernées pour limiter la divulgation des données à des publics spécifiques ; les engagements du Responsable de la sécurité des données (*DSO*) de ne pas divulguer ou utiliser les données de certaines manières ; ou les demandes par les personnes concernées de restreindre ou de limiter d'une autre manière l'utilisation ou la divulgation des données ;

f. Une description très détaillée de toute mesure prise par le Responsable de la sécurité des données (*DSO*) pour pseudonymiser les données (« Méthodes de pseudonymisation »), et de toute erreur, biais ou variance que les Méthodes de pseudonymisation peuvent introduire dans l'ensemble de données ;

g. Une évaluation initiale générale du niveau de risque que l'ensemble de données ou les Champs de données spécifiques pourraient indiquer, par référence à la Section 514 ; et

h. Toute autre information raisonnablement susceptible d'être nécessaire à un chercheur pour s'acquitter de ses obligations en vertu du présent Code. Un Livre de codes doit fournir suffisamment de détails et d'explications pour permettre aux chercheurs de s'acquitter de leurs obligations en vertu du Code, sans contenir de données à caractère personnel de l'une des personnes concernées dans l'ensemble de 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.

Nous recommandons un organe d'examen indépendant chargé d'évaluer et de hiérarchiser les projets et de limiter le nombre total de demandes à une quantité qui soit réalisable dans un délai donné. En outre, comme le recommande le rapport de l'EDMO, l'organe indépendant devrait examiner si toute recherche proposée est éthique, possible et légale. Ce faisant, de nombreuses limitations peuvent être introduites, comme indiqué dans le rapport, y compris des formes de limitation sur le champ d'application.

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 ?

Nous pensons qu'il est plus approprié de laisser cela à l'organe intermédiaire lui-même, en fonction de ses propres capacités opérationnelles, et de le communiquer par la suite aux plateformes dans un format et une périodicité convenus.

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

Nous pensons qu'il est approprié qu'un organe tiers, indépendant des plateformes et des chercheurs, surveille et évalue les demandes de recherche. L'organe doit posséder une expertise significative dans les domaines suivants :

- Les obligations légales des chercheurs et des entreprises en vertu de la législation européenne (en particulier le RGPD)
- Les méthodes de recherche
- L'éthique de recherche
- Les technologies améliorant la confidentialité

À cette fin, nous attendons de cet organe qu'il soit composé non seulement d'experts scientifiques, mais aussi d'experts juridiques, techniques et réglementaires. Nous soutenons fermement l'idée qu'un tel organe soit développé au niveau de l'UE, conformément aux recommandations du rapport de l'EDMO sur l'accès aux données.

En outre, un tel comité est particulièrement bien placé pour appuyer les importants exercices de hiérarchisation qui seront nécessaires pour identifier et développer des jeux de données qui répondent aux besoins de la plupart des chercheurs et/ou de la recherche la plus importante.

> *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 ?*

Les plateformes devraient avoir une possibilité raisonnable de s'opposer à la demande d'un chercheur et de demander une modification de celle-ci au minimum lorsque les données ne peuvent être fournies, lorsque cela constituerait une charge excessive ou une duplication des données dont disposent les chercheurs, lorsque cela porterait atteinte au secret des affaires, au droit d'auteur ou à des intérêts commerciaux légitimes. Dans l'exercice de cette possibilité raisonnable de s'opposer, les plateformes devraient disposer d'un délai raisonnable (au moins 30 jours ouvrables) pour effectuer une recherche en réponse à une demande et, par la suite, soit 1) donner suite à la demande, 2) s'opposer à la demande, 3) modifier la demande ou 4)

demander une extension limitée dans le temps pour rendre une décision, y compris une brève explication des raisons pour lesquelles l'extension est nécessaire.

Les plateformes devraient également être en mesure de répondre que les données ne sont pas disponibles après une recherche raisonnable.

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

Les projets devraient être sélectionnés sur la base de leur mérite scientifique, tel que déterminé par un examen éthique, méthodologique et juridique requis (RGPD) tel que décidé par des pairs examinateurs qualifiés. Nous ne pensons pas qu'il soit important que les équipes soient réparties entre plusieurs pays, et le rapport de l'EDMO souligne en effet qu'il peut y avoir des problèmes avec les lois contradictoires des États membres.

De par notre expérience, nous recommandons que l'organe d'examen indépendant détermine le processus approprié pour ces sélections, au moins en partie sur la base de ses capacités opérationnelles. Plus directement, en vertu du RGPD, tous les projets doivent être examinés *ex ante*, ce qui mettra à rude épreuve l'organe d'examen. Il devrait mettre au point des systèmes (y compris des frais de service pour les entreprises et les chercheurs) qui l'aideront à gérer et à répondre à la demande de manière appropriée.

> *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 ?*

Pour répondre efficacement et en temps opportun aux demandes de recherche, il devrait y avoir une cadence prévisible des demandes. Cela inclut des délais raisonnables et prévisibles pour que les chercheurs soumettent des demandes et pour que les plateformes fournissent les données en réponse à celles-ci. Cela permettra aux plateformes de disposer d'un niveau de personnel approprié pour répondre aux demandes et aux chercheurs de pouvoir compter sur les délais de réponse anticipés. La durée des projets est moins importante que le nombre de projets en cours d'exécution. Tous ces projets de recherche impliquent des exigences en matière d'ingénierie, de données et de conformité juridique, ainsi que d'autres exigences en matière de personnel, en particulier si le projet nécessite des mises à jour continues ou périodiques des données. Ces besoins en personnel signifient intrinsèquement des contraintes de personnel sur le nombre de projets concomitants qui peuvent être soutenus.

En ce qui concerne le calendrier des demandes de recherche, nous recommandons des recherches prospectives ou des demandes de données. Il peut être difficile ou impossible de collecter rétroactivement des données exactes ou de s'assurer que les données d'activité correspondent à la signification que nous leurs donnons. Ainsi, les périodes de recherches à long ou à court terme comptent moins que la nature prospective des données demandées ; il

est plus facile de collecter les bonnes données pour l'avenir pendant des périodes plus longues que d'essayer de comprendre les données rétrospectives.

> *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 ?*

Nous recommandons que la recherche soit revue afin de s'assurer qu'elle ne révèle aucune donnée à caractère personnel individuelle avant toute forme de publication (préimpression, document de travail, présentation en conférence, soumission à une revue, etc.). C'est la pratique que nous employons dans le cadre de notre partage actuel de données avec les chercheurs.

Nous recommandons également un processus d'examen par les pairs pour toutes les publications afin de s'assurer qu'elles répondent aux normes exigentes de la recherche scientifique.

> *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 ?*

Comme mentionné ci-dessus, nous effectuons actuellement, et recommandons, l'examen préalable à la publication dans un but de protection de la confidentialité. Cela ne compromet pas l'indépendance des chercheurs (il s'agit d'un examen relatif à la confidentialité, pas d'un blocage de pré-publication) et pourrait être mené par un tiers pour en assurer l'indépendance.

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 ?*
> *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 ?*

Les jeux de données devraient être adaptés à l'objectif de répondre à un groupe ou à un domaine de questions ; ils ne devraient pas être personnalisés ou spécifiquement adaptés à un seul projet de recherche. En effet, la création d'un jeu de données de haute qualité et bien documentées nécessite un investissement important de ressources, et comme indiqué à la question C.7, ces ressources sont limitées et temporairement consommées une fois allouées aux projets. Nous recommandons également fortement qu'une fonction de hiérarchisation ait lieu en dehors des entreprises, et idéalement auprès de l'intermédiaire indépendant proposé.

Cette hiérarchisation doit permettre de canaliser la communauté académique et de déterminer les priorités afin que les entreprises puissent répondre aux besoins les plus importants et fournir un maximum d'avantages à la communauté scientifique au sens large et à l'intérêt public, plutôt qu'à un petit groupe de chercheurs influents ou d'élite.

> *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?*
>
> *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 ?*

Le récent rapport de l'EDMO spécifie des méthodes pour rendre ce système pérenne, principalement en mettant en place un organe de gouvernance indépendant pour gérer les examens préalables à la recherche exigés par le RGPD. En outre, il fournit des exemples de mesures et un système d'évaluation des risques que l'organe pourrait utiliser, ainsi qu'un processus de mise à jour de ceux-ci lorsque de nouvelles technologies ou questions de recherche sont développées.

Préciser les technologies visant à améliorer la confidentialité ou d'autres méthodologies aujourd'hui limite la seule chose que nous savons aujourd'hui avec certitude au sujet de la recherche : l'avenir apportera du changement. En outre, la spécification devrait porter sur la sensibilité des données, l'identifiabilité et les préjudices de leur utilisation, et non sur des mesures spécifiques. Ensuite, il conviendrait d'appliquer les mesures qui sont actuellement à la pointe de la technologie.

> *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 ?*

Les limitations de durée/temps sont toujours une bonne idée du point de vue d'une bonne hygiène des données, et nous devrions nous attendre à ce que les chercheurs soient tenus de présenter une nouvelle demande chaque année, au minimum, pour conserver leur accès et assurer la sécurité des systèmes et des données concernés. En outre, afin de se conformer à la loi, comme indiqué dans le rapport de l'EDMO, il y aura des exigences pour limiter et minimiser l'accès aux données.

Cela dit, nous pensons au final que la meilleure façon de déterminer les limites appropriées sera de mettre en place l'organe de gouvernance indépendant susmentionné. Étant donné que les besoins des chercheurs changeront, de même que les progrès en matière de garanties, le fait de spécifier des limites à l'avance enfermera les chercheurs dans l'état actuel des choses.

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 recommandons vivement l'utilisation de salles blanches virtuelles afin que toutes les analyses de données et les accès soient à la fois déconnectés des autres sources de données et que toutes les activités soient enregistrées pour une analyse ultérieure en cas d'abus.

Nous ne nous prononçons toutefois pas sur la question de savoir qui doit gérer ces salles blanches, car il y a de bonnes raisons pour que les instituts de recherche, les gouvernements et les plateformes en soient propriétaires et exploitants. La meilleure façon de progresser est de conserver une certaine souplesse quant aux personnes concernées et de définir des normes techniques claires pour ces environnements. (par ex. tests d'intrusion, ISO 27001, etc.)

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

La manière dont les données doivent être stockées dépend de la sensibilité des données. Certaines données peuvent être transmises directement aux chercheurs ; certaines devraient être détenues par un tiers pour que les chercheurs puissent y accéder ; et certaines devraient être détenues par les plateformes pour que les chercheurs puissent y accéder. Encore une fois, nous réaffirmons l'importance d'un organe indépendant pour procéder à ces évaluations, car il s'agit essentiellement du type de mesure à mettre en œuvre.

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 ?

Nous pensons que le monde de la recherche bénéficie du développement d'un organe indépendant, comme le souligne le rapport de l'EDMO sur l'accès des chercheurs aux données. Bien que cet organe ne fasse pas partie d'une autorité de protection des données (APD) nationale ou européenne, il devrait être approuvé par ces organes afin de fonctionner dans le cadre d'un code de conduite en application de l'Article 40 du RGPD. En outre, nous prévoyons qu'un tel organe indépendant pourra consulter les APD à l'avenir au sujet des mesures.

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 ?

De par notre expérience, nous recommandons vivement de séparer l'accès aux données du financement de la recherche, car les deux tâches sont fondamentalement différentes. D'après notre expérience, il est préférable pour les chercheurs de sécuriser d'abord l'accès aux données dont ils ont besoin pour une étude particulière (par exemple par l'intermédiaire de l'organe indépendant proposé) puis d'utiliser cet octroi d'accès aux données dans le cadre de leurs efforts de collecte de fonds.

Toutefois, le soutien technique devrait être fourni par l'entité qui détient les données auxquelles les chercheurs ont accès, dans la mesure où ils en ont besoin. En outre, les plateformes devraient fournir une aide concernant les données qu'elles partagent, tant que ce soutien est habituel et raisonnable (par exemple, un délai de réponse suffisant est accordé, et les demandes de soutien peuvent être triées).

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 ?

Nous renvoyons ici au rapport de l'EDMO sur le sujet dans son ensemble. Ce rapport appelle à un organe d'examen indépendant pour confirmer que toutes les parties se sont engagées à respecter les obligations qui leur incombent en vertu du RGPD.

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 ?

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 ?

Nous soutenons la modernisation des lois qui régissent l'Internet, y compris les lois concernant la vie privée et les contenus nuisibles et illégaux en ligne. Sur les questions des contenus, nous soutenons la mise en place d'audit et de surveillance par des tiers indépendants.
Ainsi nous sommes **déjà à la pointe du secteur en matière de transparence.**

- Depuis plusieurs années, nous publions des données détaillées dans le Rapport sur l'application des Standards de la Communauté. Vous pouvez consulter le nombre de contenus supprimés, le nombre de contenus que nous détectons de manière proactive, et même le nombre de contenus négatifs que les gens voient sur nos services. Cela nous permet de voir les progrès que nous avons accomplis, les domaines dans lesquels nous devons encore travailler, et nous rend publiquement responsables de ces progrès.

- Et, parce que nous ne voulons pas que les gens nous croient sur parole, nous soumettons ce rapport de transparence à un audit indépendant réalisé par un cabinet comptable de référence.

Mais lorsqu'il s'agit de **vérifier l'approche globale de la modération de contenus**, il reste encore beaucoup de travail à faire.

C'est ce que fait le **Digital Trust and Safety Partnership (DTSP)**.

- Le DTSP a réuni des plates-formes Internet pour convenir de **bonnes pratiques** générales en matière de modération de contenus, et il s'efforce d'ajouter des détails et des spécificités qui varieront en fonction du type de services.
- Et maintenant, le DTSP **travaille sur la manière dont un mécanisme d'audit et de surveillance devrait fonctionner**. Ce sont des questions compliquées, et les détails auront leur importance.
- En 2022, les entreprises du DTSP feront l'objet **d'une sorte d'audit à blanc appelé "auto-évaluation"**.
- Puis, en 2023, les entreprises DTSP seront soumises à un **audit indépendant**.

Nous espérons qu'un **système de type DTSP deviendra une norme du secteur au niveau mondial dans les deux prochaines années**.

- Ce type de **processus fondé sur des normes et des preuves** est ce dont nous avons besoin pour savoir comment les sociétés Internet s'en sortent en matière de modération de contenus.
- Et c'est ce dont nous aurons besoin **pour tenir les plateformes responsables** lorsqu'elles n'effectuent pas un travail efficace.

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

Pour encourager le partage de données avec les chercheurs, au-delà de la limitation des données partagées avec les chercheurs au minimum nécessaire et pertinent pour atteindre les objectifs de la recherche, la réglementation devrait exclure la divulgation de secret des affaires ou d'informations commerciales confidentielles par les plateformes, y compris, mais sans s'y limiter :

- les codes source ;
- les données créées à des fins de sécurité ou d'intégrité ;
- les données dérivées générées par la plateforme couverte qui ne sont pas actuellement divulguées aux utilisateurs ;
- l'apprentissage automatique ou autres coefficients de modèles algorithmiques ;
- les données confidentielles ; et
- les données que la plateforme est contractuellement empêchée de divulguer.

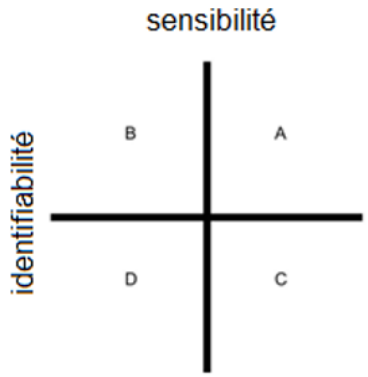
Les plateformes devraient avoir une formation interne, des politiques et des processus d'examen des ressources pour s'assurer que les diligences nécessaires soient effectuées avant que les informations ne soient partagées en externe. Inversement, les chercheurs avec lesquels

des données sont partagées devraient s'engager à empêcher toute diffusion ultérieure de données au-delà des parties à l'accord de partage et, lorsqu'il existe un accès continu, les chercheurs devraient être tenus d'attester régulièrement qu'ils continuent de respecter les mesures et contrôles de données applicables.

Les plateformes devraient également avoir la possibilité d'examiner les rapports sur la base des données qu'elles fournissent à la recherche avant la publication pour s'assurer qu'aucune donnée à caractère personnel ou information commerciale ou confidentielle n'est incluse.

Annexe - Le cadre d'évaluation des risques d'identifiabilité et de sensibilité

Ce cadre évalue le risque total inhérent à toute investigation scientifique particulière en fonction des deux attributs des données qui, selon le RGPD, sont pertinents pour un partage sûr des données : la sensibilité et l'identifiabilité des données sous-jacentes. Pour ce faire, nous mesurons la sensibilité et l'identifiabilité faibles et élevées.



L'identifiabilité et la sensibilité se combinent pour déterminer le risque de partage des données.

En combinant le niveau de sensibilité et d'identifiabilité de toute donnée, nous obtenons quatre catégories, ABCD, qui sont dans l'ordre du plus risqué au moins risqué, de sensibilité et d'identifiabilité élevées à faible. Aux fins de cet exercice, guidé par le RGPD lui-même, nous partons de l'hypothèse que la sensibilité des données comporte intrinsèquement plus de risques que l'identifiabilité des données, ce qui explique pourquoi le coin Nord Ouest est B et le coin Sud Est est C.

Puisque chacune de ces catégories représente un risque pour l'analyse de ces données, nous pouvons alors concevoir des mesures de protection qui contribuent à atténuer le niveau et le type spécifiques de chacune d'entre elles. Le tableau ci-dessous illustre la manière de faire correspondre les mesures de protection existantes aux risques spécifiques présents dans chaque catégorie :

Mesure de protection	Exemple canonique
A	1) FFRDC ou équivalent (par exemple, salle blanche physique avec contrôle) 2) Environnement virtuel avec un accès contrôlé et suppression restreinte des données, accord juridique de non redistribution ET de non reidentification.
B	1) Accès contrôlé à l'API ou au téléchargement de jeux de données, accord juridique de non redistribution et de non réidentification, audit de sécurité. 2) Environnement virtuel avec un accès contrôlé et suppression restreinte des données, accord juridique de non redistribution ET de non réidentification.
C	Accès contrôlé à l'API ou au téléchargement des jeux de données, accord juridique pour ne pas redistribuer
D	Vérification du chercheur (si requis en fonction de la source des données)

Note : La distinction entre C et B concerne l'accès équitable - garantir que certaines données soient accessibles aux chercheurs des institutions qui n'ont pas les ressources nécessaires pour se conformer aux exigences de sécurité. La décision prise est qu'il y a moins de risque dans l'identifiabilité que dans la sensibilité.

Application de ce cadre dans les projets de recherches individuels

Ce cadre s'adapte soigneusement à l'évaluation de tout projet de recherche donné, mais il nécessite deux analyses pour être correctement appliqué. Pour l'utiliser, il faut tenir compte à la fois des contributions à un projet de recherche et des "déductions de recherche" qui seront faites au cours du processus d'analyse. Les contributions sont faciles à imaginer : il s'agit des données que nous ou d'autres entreprises partageraient, qui constituent la base de l'analyse d'un projet. En revanche, les "déductions de recherche", sont aussi importantes mais moins intuitives. Étant donné que le RGPD exige une évaluation des données, et pas seulement celle que nous pourrions partager,

nous devons également examiner les données intermédiaires que les chercheurs peuvent créer lorsqu'ils mènent un projet. Tangiblement, cela signifie que pour comprendre le risque, nous devons connaître l'identifiabilité et la sensibilité de toutes les contributions, analyses intermédiaires et résultats.

Un exemple extrême permet de clarifier cette nécessité. Les chercheurs qui s'intéressent aux méthodes de réidentification peuvent commencer avec des données de type D -- non identifiables, non sensibles -- mais finir par obtenir quelque chose qui se rapproche du type A -- individuel et sensible. Dans ce cas, nous recommandons des mesures de protection adaptées au type A - le niveau de risque le plus élevé de toute l'activité de recherche.

En outre, étant donné que le RGPD exige une analyse d'équilibre de chaque activité de recherche individuelle, *ex ante*, nous pouvons estimer à la fois les contributions et les déductions, puis appliquer les mesures de protection les plus pertinentes pour les données à plus haut risque. En pratique, cela signifie que si des données de type A font partie d'une recherche, nous devons appliquer des garanties suffisantes pour A dès le début.

En faisant correspondre toutes les contributions et déductions possibles, nous obtenons 16 scénarios possibles. Le tableau ci-dessous présente ces 16 scénarios, ainsi qu'un exemple du type de projet de recherche qui pourrait vraisemblablement contenir ces contributions et déductions.

Données d'entrée	Déduction de recherche*	Seuil de mesures de protection	Exemple de proposition de recherche	Type de mesures de protection recommandées
A	A	A	Sondage individuel de l'opinion politique pour ensuite prédire le comportement politique individuel	A
A	B	A	Observation du comportement politique individuel pour construire des agrégats au niveau géographique (par exemple, sondage à la sortie des bureaux de vote).	A
A	C	A	Création d'un classificateur de contenus "manifestement publics" à partir de contenus publics diffusés sur des pages publiques.	A
A	D	A	Construire les données pour le programme de Meta *Data for Good Products* (créer des agrégats de mobilité au niveau d'un pays à partir de données de localisation au niveau individuel)	A
B	A	A	Utilisation de déductions écologiques à partir de suivis de pages publiques pour prédire le choix de vote individuel	A
B	B	B	Analyse de la qualité des hôpitaux en utilisant des résultats agrégés par condition	B
B	C	B	Utilisation de déductions écologiques à partir des pages publiques suivies pour prédire l'équipe sportive préférée d'une personne.	B
B	D	B	Développement d'une métrique de prévalence des communications climatiques à partir de données présentes sur des ensembles de mots de groupes privés	B
C	A	A	Utilisation d'une enquête sur les données de marques pour prédire la sexualité	A
C	B	B	Utilisation d'une enquête sur les données de marques pour prédire les élections locales	B
C	C	C	Utilisation d'une enquête sur les données de marques pour construire un modèle type commercial	C

C	D	C	Utilisation d'une enquête sur les données de marques pour décider où ouvrir de nouvelles coopératives d'épiceries	C
D	A	A	Recherche sur la ré-identification de l'auteur à partir de données purgées d'information personnelles, tels que les données textuelles	A
D	B	B	Prédire les taux de maladies cardiaques dans un quartier en fonction de l'emplacement et du nombre d'épiceries et de restaurants	B
D	C	C	Étude de réidentification de Netflix	C
D	D	D	Toute recherche observationnelle anonyme et agrégée	D
*Ce n'est PAS ce qui est publié, mais ce qui est produit au cours de la recherche. Par exemple, cela inclut les déductions générées par le chercheur, qu'elles soient ou non partagées par le chercheur.				

Comme illustré dans le tableau, les mesures de protection pour les données à plus haut risque sont appliquées dans tous les cas, ce qui signifie que sur les 16, il y a sept scénarios qui nécessitent des mesures de protection de type A, cinq de type B, trois de type C et un de type D.

L'utilisation de ce cadre permet de lancer la création d'un précédent; ce n'est pas une formule fixe

Enfin, ce cadre est considéré à juste titre comme un moyen d'initier la création d'un précédent afin de créer *ex ante* un organisme décisionnel. Il s'agit ici d'aider cet organisme à prendre les bonnes décisions dès le début de sa mise en œuvre et d'intégrer des lignes directrices dans son fonctionnement. Il ne s'agit absolument pas de s'en tenir à l'avenir uniquement à cette analyse ou mesures de protection.

En présentant ces catégories et mesures de protection, nous espérons accomplir deux choses pour ces futurs examinateurs *ex ante*. Premièrement, nous transformons toute future analyse de risque en une évaluation relative au lieu d'une évaluation absolue. Les examinateurs peuvent se demander : "*l'activité proposée est-elle plus ou moins risquée par rapport à l'un de ces exemples* ?"

Deuxièmement, nous évitons de verrouiller des technologies de confidentialité particulières. Notez, par exemple, que la confidentialité différentielle n'apparaît nulle part dans ce document jusqu'à présent. Ceci est dans l'esprit de créer un précédent. Les données privées différentielles seraient peu identifiables, et donc automatiquement un B ou un D. Si à l'avenir la confidentialité différentielle est brisée, les futures analyses *ex ante* évalueraient ces données comme étant de type A ou C.