Arcom

Le régulateur de la communication audiovisuelle et numérique

Accès des chercheurs aux données des plateformes : synthèse des réponses à la consultation de l'Arcom et propositions

Juin 2023

Résumé exécutif

En mai 2022, l'Arcom a publié une consultation publique sur l'accès aux données des plateformes en ligne à des fins de recherche. Cette démarche a été initiée pour contribuer à la mise en œuvre du Règlement sur les Services Numériques (RSN, ou Digital Services Act en anglais). En effet, les conséquences sociétales induites par les usages en ligne font désormais partie des prérogatives des régulateurs nationaux et européens, sur la base du constat des **limites évidentes** d'un **modèle d'auto-régulation des plateformes.**

Le Règlement sur les Services Numériques est ainsi venu redéfinir les responsabilités des plus grandes plateformes et moteurs de recherche. Dans son article 40, le texte crée un mécanisme innovant ancrant l'accès des chercheurs aux données de ces opérateurs. Dans les faits, le succès et la durabilité du RSN reposeront en partie sur **le travail mené par les chercheurs**, qui viendra **éclairer le débat public** et **nourrir la régulation.**

L'Arcom a proposé à différents types d'acteurs issus du monde de la recherche, de la société civile, mais également du secteur privé de s'exprimer sur cette question complexe. La consultation se structurait autour de cinq thèmes de réflexion **(1) partage d'expériences d'utilisations de données de ces services, (2) gouvernance, (3) construction des projets scientifiques, (4) protection des données et considérations techniques, et (5) faisabilité des accès et incitations.**

L'analyse des réponses à la consultation publique a eu pour objectif d'identifier les principales difficultés rencontrées par les chercheurs et d'envisager comment les régulateurs nationaux et européens pourraient faciliter la bonne mise en œuvre du nouveau règlement, mais également comment ils pourraient s'appuyer sur les travaux de recherche.

Enseignements principaux de la consultation

Les entretiens menés et les réponses reçues dans le cadre de la consultation ont permis de dégager quatre points de consensus sur la situation actuelle d'accès aux données des plateformes pour les chercheurs. Ces points, qui ont déjà été intégrés en partie dans la refonte du cadre règlementaire, ont permis de nourrir une série de propositions visant non-seulement à faciliter la mise en place des nouveaux accès mais également à contribuer à la vitalité de l'écosystème de recherche.

#1 : La situation pré-RSN en matière d'accès aux données des plateformes est sous-optimale.
En l'absence d'un cadre légal dédié, les accès ont jusqu'à aujourd'hui été majoritairement permis par les plateformes de manière volontaire, concentrant de fait les recherches sur les services les plus allants en la matière. Cette hétérogénéité dans les accès permis par les plateformes a pour effet d'influencer la conception des questions de recherche et des méthodologies mises en œuvre et génère ainsi des variations substantielles entre les analyses qu'il est possible de mener sur les différentes plateformes.

#2 : Les modalités d'accès aux données à des fins de recherche doivent être adaptées à la sensibilité des données concernées.
Les données jugées sensibles ou à risque doivent être protégées. Leur accès devra être encadré selon des modalités garantissant sécurité, proportionnalité et confidentialité, dans le respect du règlement général sur la protection des données (RGPD). A l'inverse, les données publiques ou agrégées devraient pour leur part être mises à disposition du

grand public, en adoptant par exemple les principes de l'open data, qui permet la libre mise en place d'initiatives par la société civile.

#3 : Les chercheurs sont demandeurs de mécanismes clairs, efficaces et transparents

Les chercheurs formulent la demande de processus standardisés, transparents et clairs dans leurs demandes d'accès aux données. Une majorité des répondants soulignent l'utilité qu'aurait l'instauration d'un intermédiaire entre chercheurs et plateformes, qui se fonderait sur une importante expertise technique. Il sera toutefois important de garantir la transparence et l'indépendance d'un tel intermédiaire.

#4 : Il est important de capitaliser sur les expertises existantes afin de contribuer à la vitalité de l'écosystème de recherche

Il existe des expertises importantes en matière de partage et de traitement de données, sur lesquelles il sera crucial de s'appuyer afin de permettre la bonne mise en œuvre des nouvelles modalités d'accès, et de faciliter leur appropriation par le plus grand nombre de chercheurs possible. La France dispose de laboratoires de recherche produisant des travaux à la pointe de l'état de l'art en matière d'études des plateformes dont le rôle dans la transmission de savoir et de techniques devra être pérennisé. Enfin, la CNIL, de par son mandat, dispose d'une grande expertise en matière de protection des données personnelles. Elle aura un rôle particulièrement important afin d'assister l'Arcom dans la montée en puissance des accès aux données.

Propositions de l'Arcom suite à la consultation

La connaissance produite par la communauté de recherche a un rôle stratégique : nourrir la régulation, et éclairer le débat public. Ce document met en avant un certain nombre de propositions concrètes de l'Arcom visant à permettre une collaboration efficace et mutuellement bénéfique entre régulateurs et chercheurs. Les nombreux échanges avec la communauté de recherche ont d'ailleurs déjà servi de base à l'Arcom pour contribuer[1] aux travaux de la Commission européenne portant sur l'acte délégué qui précisera les conditions des accès aux données prévus par l'article 40 du RSN.

(1) Contribution à l'animation et à la vitalité de l'écosystème de recherche en France

a) Maximiser les synergies au sein de l'écosystème

- Faciliter le partage d'expérience entre chercheurs
- Optimiser la réutilisabilité des jeux de données
- Valoriser l'implication et la montée en compétence de jeunes chercheurs
- Valoriser le déploiement d'outils computationnels publics
- Soutenir les approches pluridisciplinaires et transnationales
- Soutenir le déploiement de solutions techniques sécurisées
- Favoriser une logique d'open data s'agissant des données les moins sensibles

b) Réguler en s'appuyant sur les travaux de recherche

- Contribuer à la visibilité des chercheurs et de leurs travaux dans le débat public
- Identifier les problématiques de recherche émergentes
- Développer des capacités d'analyse agiles
- Faciliter la hiérarchisation des besoins en nouvelles données et de mise à jour des API

[1] Réponse de l'Arcom à l'appel à contribution de la Commission européenne sur l'acte délégué complétant l'article 40 du RSN - Feedback from: Arcom (europa.eu).

- Protéger les accès existants et inciter la mise en place d'API
- Défendre un accès aux données publiques pour les médias, vérificateurs d'informations et pour les associations de la société civile
- Engager une réflexion autour des dons de données et du *scraping*

(2) Contribution de l'Arcom à l'efficacité des mécanismes d'accès aux données mis en place par le RSN et le Code européen renforcé de bonnes pratiques contre la Désinformation

a) Assurer la bonne circulation de l'information au niveau français

- Expliquer et promouvoir les nouveaux mécanismes d'accès aux données
- Fournir exceptionnellement un accompagnement à certains projets
- Œuvrer à la transparence des processus d'accès

b) S'inscrire dans une logique de « réseau de régulateurs »

- Faciliter le travail d'agrément du CSN d'établissement des VLOPSEs
- Opérer la liaison entre le niveau national et européen
- Suivre les travaux de constitution d'un tiers indépendant

Table des matières

Résumé exécutif 3

1. Introduction 7

2. L'accès aux données des plateformes avant le RSN 8

a) L'importance cruciale de l'accès aux données des plateformes 8

b) Besoins des chercheurs en données et difficultés d'accès 9

Cartographie des approches méthodologiques et de leurs besoins en données 9

Difficultés d'accès aux données 13

3. Périmètre de la régulation et construction d'une nouvelle gouvernance 17

a) Evolution du cadre règlementaire européen 17

Le Règlement sur les Services Numériques 18

Nouveau code de bonnes pratiques contre la désinformation 24

b) Construction d'une nouvelle gouvernance 29

Sous-optimalité de la situation actuelle 29

Pertinence d'un accès aux données basé sur le risque 29

Le besoin de mécanismes clairs, efficaces et transparents 30

Capitaliser sur les expertises existantes pour contribuer à la vitalité de l'écosystème de recherche 31

4. Propositions 32

a) Contribuer à la vitalité de l'écosystème de recherche français 33

a.1) Maximiser les synergies au sein de l'écosystème 33

a.2) Réguler en s'appuyant sur les travaux de recherche 36

b) Tirer pleinement parti des nouvelles possibilités d'accès aux données 38

b.1) Assurer la bonne circulation de l'information au niveau français 38

b.2) S'inscrire dans une logique de « réseau de régulateurs » 39

Références 42

1. Introduction

En mai 2022, l'Arcom a publié une **consultation publique** sur l'accès aux données des plateformes en ligne à des fins de recherche. Cette démarche a été initiée dans le but de nourrir les réflexions déjà existantes sur le sujet, notamment au niveau européen et d'enrichir les réflexions de l'Arcom sur le rôle qu'elle devra jouer pour faciliter ce **meilleur accès aux données** pour le monde de la recherche.

Les réseaux sociaux, les plateformes de partage de vidéos et les moteurs de recherche redéfinissent la façon dont les contenus, notamment d'information, sont consommés et partagés. Au cours des dernières années, les opportunités multiples offertes par les plateformes ont ouvert de nouveaux espaces d'expression citoyenne et de circulation de l'information. Néanmoins, à ces aspects positifs se sont ajoutées les dérives des nouvelles formes de sociabilité et d'expression des individus en ligne, tels que les phénomènes de manipulation de l'information ou de haine en ligne.

De nombreux acteurs académiques et issus de la société civile ont régulièrement exprimé le besoin d'étudier les phénomènes en œuvre sur les plateformes à partir de données provenant des opérateurs de services eux-mêmes. Ces acteurs ont notamment contribué à mettre en lumière l'opacité des choix opérés par les plateformes en matière algorithmique, de politique de modération des contenus, ou de mise à disposition de ressources pour le grand public ou les communautés de recherche.

Le questionnaire était destiné à différents types d'acteurs issus du monde de la recherche, de la société civile, mais également du secteur privé. Il était structuré autour de 5 thèmes : **(1) partage d'expériences d'utilisations de données de ces services, (2) gouvernance, (3) construction des projets scientifiques, (4) protection des données et considérations techniques, et (5) faisabilité des accès et incitations.**

L'Arcom a reçu 15 réponses entre mai et novembre 2022, la plupart venues de consortiums de chercheurs ou de laboratoires de recherche, mais aussi de la part d'organisations privées, de représentants de la société civile, ou de certaines plateformes elles-mêmes.[2] La diversité des profils des répondants a permis de porter un regard complet sur les différents enjeux que soulève la question de l'accès aux données. A ces réponses se sont ajoutés de nombreux échanges avec différentes parties prenantes, qui ont également nourri la réflexion de l'Arcom.

L'initiative de la consultation remonte à l'automne 2021, période coïncidant avec les négociations du futur Règlement sur les Services Numériques (RSN) (ou *Digital Services Act* en anglais). L'analyse des contributions a donc été menée dans un contexte d'évolution profonde du cadre réglementaire européen. Le Règlement sur les Services Numériques[3] vise à mieux responsabiliser les plus grandes plateformes et moteurs de recherche. Ce texte met notamment en place par son article 40 un mécanisme innovant ancrant l'accès des chercheurs aux données de ces opérateurs. Au RSN s'ajoutent le nouveau *Code européen renforcé de bonnes pratiques contre la désinformation du 16 juin 2022 (ci-après « Code de bonnes pratiques contre la désinformation »)* signé par plusieurs parties prenantes, et les efforts entrepris par l'EDMO (Observatoire Européen des Médias Numériques)[4] pour mettre en place un cadre d'accès des

[2] Ces contributions sont disponibles en intégralité en annexe de cette synthèse.
[3] La première proposition du futur *Règlement sur les Services Numériques* a été présentée en décembre 2020 par la Commission européenne. Le Conseil de l'Union européenne a quant à lui adopté son orientation générale fin 2021, et le Parlement Européen a présenté ses propositions d'amendement en janvier 2022. Le RSN a été publié le 27 octobre 2022 au journal officiel de l'UE : EUR-Lex - 32022R2065 - EN - EUR-Lex (europa.eu).
[4] Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access – EDMO, 2022.

chercheurs aux données des plateformes dont le traitement peut présenter des « risques pour les droits et les libertés des utilisateurs ».[5]

Ce document propose une synthèse des réponses à la consultation en les mettant en perspective avec les évolutions récentes du cadre règlementaire. Il fixe également une feuille de route sur le rôle que pourra jouer l'Arcom pour assurer la pleine effectivité et efficacité du nouveau cadre d'accès aux données.

2. L'accès aux données des plateformes avant le RSN

a) L'importance cruciale de l'accès aux données des plateformes

Les nouvelles problématiques soulevées par le développement des réseaux sociaux, des plateformes de partage de vidéos en ligne et des moteurs de recherche illustrent que la question est devenue un enjeu public. En effet, l'environnement informationnel actuel ne se définit plus par l'addition de secteurs dont les frontières seraient hermétiques : audiovisuel et numérique ; médias édtiorialisés (télévision, radio, presse) et nouveaux services de consommation de contenus (réseaux sociaux, applications) ; modes de réception historiques et terminaux de demain ; médias nationaux, européens et internationaux. Les recoupements sont au contraire désormais de plus en plus importants.

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

Les usages sur internet rivalisent à présent avec ceux des médias traditionnels et l'on assiste à des phénomènes de redistribution des temps d'attention consacrés aux médias et des sources choisies, qui renforcent le rôle structurant et croissant d'internet dans l'accès à l'information. Dans une étude publiée en novembre 2022 par l'Ofcom, intitulée « Media Plurality and Online News »[6], le régulateur britannique met en lumière les différents effets que peuvent avoir les opérateurs de médias en ligne sur les modes d'accès à l'information (pluralisme, connaissance de l'actualité, biais algorithmiques, polarisation, confiance dans les médias…). Ces nombreux effets ont également été documentés par de nombreux chercheurs au cours des dernières années.[7] Cette synthèse n'a pas pour but de proposer une revue exhaustive de la littérature sur les effets des usages en ligne sur nos sociétés, mais s'appuie sur les recherches existantes et souligne la nécessité de déployer de nouvelles initiatives à plus large échelle couvrant toutes les dimensions des transformations des environnements informationnels et des nouveaux risques qui leur sont associés.

[5] Voir le cadre d'évaluation des risques présenté dans le rapport du groupe de travail de l'EDMO sur l'accès aux données (EDMO, 2022 :55).
[6] Voir sur le site de l'Ofcom : https://www.ofcom.org.uk/research-and-data/multi-sector-research/media-plurality
[7] Voir par exemple : Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, *110*(3), 629-676. ; Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, *12*, 415-438. ; Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). *Social media and xenophobia: evidence from Russia* (No. w26567). National Bureau of Economic Research. ; Levy, R. E. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, *111*(3), 831-870. ; Fujiwara, T., Müller, K., & Schwarz, C. (2021). *The effect of social media on elections: Evidence from the United States* (No. w28849). National Bureau of Economic Research.

Ces nouveaux risques sont autant d'enjeux dont les pouvoirs publics doivent être en mesure de se saisir. En effet, le **modèle d'auto-régulation des plateformes a montré ses limites** et le RSN a donné de nouvelles prérogatives aux régulateurs nationaux et à la Commission européenne.

En matière de lutte contre la manipulation de l'information par exemple, la loi du 22 décembre 2018 impose aux plateformes dépassant un seuil d'audience de 5 millions de visiteurs uniques mensuels en France, de rendre accessible et visible un dispositif de signalement des contenus relatifs à la manipulation de l'information, et de déployer des mesures complémentaires en matière de transparence de leurs actions visant à lutter contre ces phénomènes. Sur ce point, malgré des améliorations notables depuis sa précédente évaluation, l'Arcom a déploré dans son dernier bilan[8] des lacunes répétées dans la publication par certains opérateurs d'informations et de données chiffrées exigées par le cadre juridique français. La mise en œuvre de cette régulation montre ainsi les limites d'un cadre essentiellement souple qui ne permet pas d'exiger des plateformes qu'elles rendent des comptes, d'une part, sur la réalité des phénomènes de manipulation de l'information sur leur service et, d'autre part, sur l'ensemble des mesures prises pour les atténuer.

Le monde de la recherche est appelé à jouer un rôle central dans la compréhension de ces phénomènes complexes, afin de permettre à la société civile de se saisir de ces enjeux et responsabiliser les plateformes, mais également afin de nourrir l'action des pouvoirs publics. Au-delà de ce que les plateformes rendent disponible, ou des analyses conduites en interne qu'elles publient, le monde de la recherche doit donc pouvoir également accéder à des données riches, fiables et appropriées selon des modalités qui ne soient pas définies par les plateformes seules.

b) Besoins des chercheurs en données et difficultés d'accès

Un large éventail d'acteurs du monde de la recherche, pouvant appartenir ou non à des structures dédiées à l'étude des plateformes en ligne, s'est développé ces dernières années. Ces structures sont aussi bien affiliées au monde universitaire qu'à des associations ou à des organes publics. Les plateformes elles-mêmes ont aussi mis en place des groupes de recherche, qui travaillent aussi bien en interne qu'en partenariat avec des chercheurs extérieurs.

Les plateformes et leurs données suscitent l'intérêt de chercheurs de disciplines variées (économie, sociologie, droit, géopolitique, informatique, psychologie, philosophie …). Ceux-ci adoptent aussi bien des méthodes dites « traditionnelles » que des méthodes nouvelles, dont certaines sont même parfois développées spécifiquement pour ce champ d'étude. Ces différentes méthodes présentent des besoins variés en termes de données, qui impliquent plusieurs techniques de collecte faisant face à des obstacles spécifiques (Institute for Strategic Dialogue, 2022 :19).

Cartographie des approches méthodologiques et de leurs besoins en données

La classification retenue ici s'inspire des catégories méthodologiques proposées par l'ISD (ISD, 2022) et mises en avant dans leur réponse à la consultation. Cette classification contient aussi bien des méthodes quantitatives que qualitatives.

La **recherche systématique**, première grande catégorie méthodologique, consiste en l'utilisation de techniques computationnelles afin d'analyser de larges bases de données.

[8] https://www.arcom.fr/nos-ressources/etudes-et-donnees/mediatheque/lutte-contre-la-manipulation-de-linformation-sur-les-plateformes-en-ligne-bilan-2021

Les protocoles de recherche systématiques se prêtent particulièrement bien à l'analyse des importantes quantités de données accumulées par les plateformes. Le champ d'applications potentielles de ce type de méthodes est extrêmement vaste, étant intrinsèquement lié aux types de données mis à disposition par les plateformes ou collectables par le biais du *web scraping.*[9] Certains protocoles de recherche ont recours à des réseaux de robots (ou *bots*) simulant le comportement d'utilisateurs réels et *scrapant* les données retournées par les plateformes. Dans les cas où ce sont les systèmes de la plateforme qui sont étudiés, le protocole de recherche consiste souvent à « duper » cette dernière dans une certaine mesure afin d'observer le fonctionnement de ses systèmes en conditions réelles.

Les méthodes de ***crowdsourcing*** portent directement le nom des techniques de collectes de données sur lesquelles elles s'appuient. Une de leurs principales applications consiste en l'analyse du comportement des utilisateurs et l'analyse des contenus leurs étant recommandés par les plateformes. Les données peuvent être collectées de manière collaborative à partir d'un panel de participants volontaires, par exemple par le biais d'extensions de navigateur (*plug-ins*)[10].

Les méthodes d'**ethnographie** étaient utilisées en sciences sociales bien avant l'arrivée d'internet. Elles présentent un intérêt particulier dans l'étude et la compréhension des communautés en ligne. Le chercheur va pour cela intégrer une communauté, suivre et éventuellement, participer à ses échanges. Ce type de méthodes permet d'accéder à des contenus difficilement accessibles comme ceux des groupes privés ou des chaines de messagerie.

Enfin, la **recherche documentaire** consiste en l'analyse de documents –principalement textuels – permettant de mieux comprendre les activités des plateformes (document légaux, financiers, rapports de transparence, conditions générales, déclaration des dirigeants, *process* internes…). Ce type de projets de recherche peut venir compléter les méthodes mentionnées ci-dessus ou bien être mené de manière autonome.

[9] Le *web scraping* consiste en l'extraction de données à partir de pages internet et à leur enregistrement de manière ordonnée dans un fichier, notamment en vue de leur analyse. Cette pratique est fréquemment utilisée en science des données. Voir aussi l'encadré 1.

[10] Un exemple étant les outils créés par l'association *AI Forensics*, qui une fois installés volontairement par des internautes, permettent à ces derniers de partager des données sur les recommandations qui leurs sont faites. Ces données sont ensuite agrégées et analysées par l'association.

Tableau 1 - Cartographie des méthodes et de leurs besoins en données

Méthodologie	Applications potentielles	Données nécessaires	Méthodes d'accès
Recherche systématique	Recherche à grande échelle sur la diffusion des contenus, sur la modération ; analyse du discours ; analyse de réseaux ; audits algorithmiques.	Grandes bases de données portant sur des métriques spécifiques (engagement, *reach*, modération) ; données d'entrée/sortie des systèmes automatisés ; données sur les interactions entre comptes	API; outils computationnels fournis par les plateformes (ex : *Crowdtangle*) ou par des tiers; *scraping* (notamment par réseau de bots); bases de données fournis directement par les plateformes (et éventuellement protégés)
Crowdsourcing	Analyse et comparaison des recommandations de contenus automatisées (audits algorithmiques notamment) ; mesures d'audience ; compréhension des comportements en ligne	Points de données collectés par plusieurs utilisateurs/entités, puis agrégés	Crowdsourcing Sondages ; collecte manuelle puis agrégation
Ethnographie	Compréhension riche des dynamiques communautaires difficilement possible en utilisant d'autres approches	Descriptions détaillées de l'expérience du chercheur au sein d'un espace numérique et/ou d'une communauté (parfois privé) ; historique de certains contenus	Collecte manuelle
Recherche documentaire	Analyse du discours des plateformes ; analyse financière ; analyse légale ; compréhension des politiques et des mécanismes de causalité derrière certaines observations techniques	Documents légaux, financiers des plateformes ; rapports de transparence ; conditions générales d'utilisation ; déclarations et communications des dirigeants ; documents internes ; documents textuels expliquant les paramètres de certains systèmes	Collecte manuelle ou automatisée

Encadré 1
Accès aux données : Méthodes coopératives et méthodes ''adversarielles''

Il est possible d'établir une distinction entre méthodes coopératives et méthodes dites *adversarielles* d'accès aux données à des fins de recherche. Les **méthodes coopératives s'appuient sur des accès permis et facilités par les plateformes** (y compris dans le cadre d'obligations légales), notamment par le biais d'API ou d'outils propriétaires. Les **méthodes adversarielles**, telles que le *scraping* ou le *crowdsourcing*, visent pour leur part à contourner les limites et les obstacles introduits par les plateformes. Dans les faits, des chercheurs souhaitant auditer une plateforme peu accommodante en matière de partage de ses données (e.g. ne disposant pas d'API) n'auront pas d'autre choix aujourd'hui que d'utiliser des méthodes adversarielles.

A terme, une amélioration de l'accès aux API permettrait donc de réduire le recours des chercheurs au *scraping*, sans toutefois l'éliminer complètement (voir ci-dessous).

Si les méthodes adversarielles permettent dans une certaine mesure de pallier aux limites ou même à la non-existence d'accès coopératifs, **ces deux méthodes ne sont pas parfaitement substituables, et sont même plutôt présentées comme complémentaires par certains chercheurs**. En effet, certaines données sont uniquement accessibles par le biais d'accès coopératifs (telles les données sur la diffusion de certains contenus). En revanche, même dans le cas hypothétique d'un progrès important en matière d'accès coopératif aux données des plateformes, les recherches basées sur des accès adversariels resteraient nécessaires selon certains chercheurs afin de ne pas dépendre uniquement des accès fournis par les plateformes, et de pouvoir conduire des audits indépendants.

Le cas du *scraping*

Le *scraping* consiste en **l'extraction automatisée et systématique de données disponibles en accès libre à partir d'un site web** (en l'occurrence une plateforme) **par le biais d'un programme tiers**. Les chercheurs ont fréquemment recours aux outils de *scraping* pour collecter des données, notamment en réponse aux insuffisances des API et logiciels créés par les plateformes, ou bien s'ils se voient refuser l'accès à ceux-ci. En effet, le *scraping* est actuellement la seule solution pour accéder aux données des plateformes en temps réel lorsqu'une plateforme ne met pas d'API à disposition des chercheurs.

Le scraping est souvent purement et simplement interdit dans les CGU des plateformes, qui se réservent ainsi le droit d'agir contre ceux qui le pratiquent. Aussi, les chercheurs procèdent généralement à l'insu des plateformes ciblées. Il arrive que certains chercheurs voient leurs accès à une plateforme révoqués suite à l'utilisation de cette méthode.[11]

[11] Voir par exemple le cas de Laura Edelson, chercheuse à l'université de New York, qui travaillait avec ses collègues sur des données publiques relatives aux publicités politiques sur Facebook : Platforms vs. PhDs: Facebook seeks shutdown of NYU research - Protocol

Difficultés d'accès aux données

En l'absence d'un cadre légal contraignant, les **accès permis par les plateformes l'ont essentiellement été de manière volontaire jusqu'aujourd'hui**, concentrant de fait les recherches sur les services les plus allants en la matière. S'il faut saluer ces initiatives, force est de constater que les recherches se sont de ce fait **surtout concentrées sur Twitter**, qui offrait différentes API dont une dédiée à la recherche. Cette ouverture a permis à de nombreux projets de voir le jour. D'autres réseaux sociaux ou moteurs de recherche font le choix d'une politique d'accès à leurs données plus restrictive, y compris pour les chercheurs.

Cette asymétrie des accès entre les différentes plateformes est problématique dans la mesure où les connaissances acquises quant au fonctionnement d'un service ne sont pas systématiquement généralisables. En effet, chaque plateforme dispose de caractéristiques propres et attire différents profils d'utilisateurs, ce qui implique que chaque plateforme peut présenter des phénomènes (et des risques) qui lui sont spécifiques.

Même les plateformes ayant des politiques d'accès plus permissives conservent actuellement un certain nombre de prérogatives lors de l'ouverture de leurs données, notamment en **imposant des conditions strictes d'utilisation de leurs API**.[12] [13] Ces prérogatives peuvent, selon certains chercheurs, compromettre les projets de recherche. Les plateformes ont notamment tendance à demander un droit de regard sur les projets, aussi bien au moment de leur formulation que préalablement à leur publication. Bien qu'un tel droit de regard soit justifié par les plateformes comme résultant de raisons de sécurité ou de respect des données personnelles, il est critiqué par certains chercheurs qui expliquent qu'il compromet leur indépendance. En parallèle, les plateformes ont également tendance à fixer des limites strictes en matière de conservation et de partage des données par les chercheurs, permettant notamment de s'assurer que des contenus qui auraient été supprimés de la plateforme soient également supprimés par les chercheurs. Les chercheurs font valoir que de telles demandes peuvent parfois être antagonistes aux processus permettant la transparence et la réplicabilité de la recherche. Enfin, rien ne garantit actuellement la **continuité des accès**, dans la mesure où les plateformes peuvent décider unilatéralement de mettre fin à ces derniers.[14]

Les chercheurs répondants à la consultation de l'Arcom ont largement confirmé faire face à ces **disparités d'accès entre les plateformes et à un manque de transparence** quant aux décisions de ces dernières, expliquant qu'elles pouvaient profondément **affecter la conception et la mise en œuvre des projets de recherche**. Plus concrètement, les chercheurs déplorent un manque de dialogue et de coopération avec les plateformes, qui s'illustre notamment par un déficit de transparence sur les données disponibles et leur représentativité, ou une absence de justification de certaines décisions des plateformes. Les chercheurs expliquent devoir parfois se résoudre à l'abandon pur et simple de certains projets.

[12] Un bilan des accès existants aux API des plateformes est annexé à la présente synthèse.
[13] Pour une cartographie des clauses posées par les plateformes lors de l'ouverture de leurs données aux chercheurs, voir Lurie, 2023 : Comparing Platform Research API Requirements (techpolicy.press)
[14] Là aussi, l'exemple de Twitter est particulièrement instructif, dans la mesure où l'entreprise a récemment indiqué son intention de changer sa politique d'ouverture de ses données, par une annonce faite sur la plateforme elle-même en février 2023. En déclarant mettre fin à une immense majorité des accès gratuits à son API, y compris pour les chercheurs, Twitter fait planer un doute sur la faisabilité de nombreux projets de recherche en cours et futurs.

Les chercheurs mettent en avant certaines **limites techniques des outils mis à disposition des chercheurs par les plateformes,** voire la **non-existence de tels outils** pour certaines plateformes qui en auraient pourtant **les moyens techniques et financiers**. Les répondants ont notamment soulevé le problème des plafonds stricts en matière de volume de données accessibles par le biais des API (*rate-limiting*), qui peuvent causer des problèmes de représentativité et de partialité des données collectées. De manière similaire, il arrive que les données accessibles par le biais d'API ne soient pas mises à disposition en temps réel, ce qui pose des questions de temporalité pour les chercheurs. Les chercheurs expliquent qu'il arrive fréquemment que la documentation mise à leur disposition par les plateformes ne soit pas à jour.

La consultation fait aussi émerger le besoin global d'une amélioration des **fonctions de recherche**, par exemple en permettant aux chercheurs de suivre la **diffusion de photos, vidéos ou URL spécifiques**, ainsi qu'en ayant accès aux données de diffusion de contenus (*reach*) par pays et sur un temps donné.

Une autre critique récurrente cible le **manque de clarté** des outils et de la présentation des résultats fournis par ceux-ci, les chercheurs répondants estimant les mises à jour trop peu fréquentes. Enfin, certains répondants mettent en avant un **besoin d'harmonisation** entre les outils et les formats des données fournis par les plateformes, qui réduirait les délais d'adaptation et permettrait une meilleure comparaison des plateformes.

Ces limites techniques intrinsèques aux outils **pourraient se multiplier** à l'avenir au fur et à mesure du **déploiement de nouvelles fonctionnalités** par les plateformes. En effet, en l'absence d'évolution parallèle des données partagées, les nouvelles fonctionnalités pourraient rendre obsolètes les accès existants. On pense par exemple au déploiement des technologies immersives (AR et VR), à la généralisation d'outils basés sur l'intelligence artificielle, ou encore au recours à des réseaux décentralisés, qui pourraient créer de nouveaux obstacles à la compréhension des services en ligne par les chercheurs s'ils ne sont pas accompagnés par le partage de nouvelles données (ISD, 2022).

Les plateformes, quant à elles, soulignent vouloir faire davantage en matière de partage de données, mettant notamment en avant l'élargissement des accès ces dernières années via la mise en place de nouveaux programmes destinés aux chercheurs. Les opérateurs soulignent toutefois la nécessité de protéger la confidentialité de certaines données (notamment les données personnelles et celles tenant au secret des affaires), ainsi que les coûts humains et financiers associés au développement et au maintien d'outils facilitant l'accès. Les plateformes veulent également éviter que leurs données soient exploitées à des fins lucratives sous-couvert d'un accès justifié par un impératif de recherche. Les plateformes déplorent enfin l'absence d'un cadre légal clair, qui permettrait notamment de définir la responsabilité juridique des différents acteurs impliqués dans le partage des données.

Encadré 2
Un débat sur la sensibilité de certaines données

Il est intéressant de noter que l'inaccessibilité de certaines données, présentée par certains chercheurs comme résultant d'un manque de volonté des opérateurs, sera à l'inverse justifiée par les opérateurs comme étant lié à la sensibilité de certaines données.

Ces cas spécifiques concernent souvent des données disponibles publiquement mais non-agrégées de manière systématique par les plateformes. Les chercheurs regrettent par exemple un manque de **données quant à l'apposition de certains labels ou compléments d'information par les plateformes sur certains contenus** d'utilisateurs (par exemple les compléments d'information en lien avec la pandémie, ou les labels liés à la présence de fausses informations). Un autre exemple concerne l'accès des chercheurs aux **données sur la mise en avant** de certains contenus par les plateformes.

Difficultés d'accès liées à la sensibilité de certaines données

Les plateformes invoquent trois grands principes afin de définir le niveau de sensibilité de leurs différents types de données (« sensibilité » n'est pas seulement entendu ici au sens du RGPD), et de justifier leurs politiques d'accès aux chercheurs :

(1) La protection des données personnelles de leurs utilisateurs, notamment dans le cadre du RGPD ;
(2) La protection du secret des affaires ;
(3) La lutte contre la manipulation de leurs systèmes par des entités ou des utilisateurs malveillants.

De nombreux types de données présentant un intérêt pour les chercheurs sont ainsi indisponibles car considérés *sensibles* par les plateformes sur la base de ces trois critères. C'est par exemple le cas de l'accès des chercheurs aux contenus échangés dans les groupes et les chaines de messagerie privés, aux paramètres gouvernant les systèmes de recommandation et de modération, voire aux algorithmes eux-mêmes, aux contenus temporaires (e.g. « *stories* » ou « *lives* » sans retransmission), aux contenus faisant l'objet de modération (contenus *downrankés*, labellisés, démonétisés ou supprimés) ou encore aux données sur les contenus publicitaires, bien que certaines plateformes aient choisi d'être plus transparentes sur ces dernières (notamment les publicités à caractère politique). Les données mises à disposition sont aussi souvent **circonscrites à un laps de temps** établi par rapport au moment de la requête (inférieur à 15 jours pour certaines plateformes) afin d'éviter le stockage par les chercheurs de données supprimées par les utilisateurs, ce qui entrave les possibilités de recherche sur des phénomènes historiques passés (par exemple les campagnes de désinformation).

Encadré 3
Protection des données personnelles & Accès des chercheurs aux données

La protection des données personnelles, strictement encadrée par la loi, est un enjeu essentiel afin de comprendre les difficultés rencontrées par les chercheurs dans l'accès aux données des plateformes.

Application du RGPD au partage de données à des fins de recherche

Le RGPD, entré en vigueur en 2018, est venu apporter un cadre règlementaire strict en matière de protection des données personnelles. Ces dernières années, plusieurs plateformes se sont appuyées sur ce texte pour justifier leur refus de partager certaines données avec les chercheurs au nom du principe de précaution (Vermeulen, 2022 :6). Les plateformes mettent régulièrement en avant le manque de clarté du RGPD concernant cette problématique.

Une difficulté notable dans l'examen du niveau de confidentialité d'un contenu donné sur une plateforme consiste en l'évaluation du **niveau de confidentialité attendu par les utilisateurs** ayant posté ledit contenu[15]. Par exemple, on peut supposer qu'un utilisateur postant sur Facebook un contenu sur son profil public ne s'attend pas forcément à ce que ce contenu circule de façon large, et ce même si ce contenu est disponible publiquement.

En réponse à cette situation, l'EDMO (Observatoire Européen des Médias Numériques), composé aussi bien d'acteurs de la société civile, de chercheurs que de plateformes, a formé un groupe de travail chargé de travailler à un code de conduite sur le sujet, tel qu'encouragé par l'article 40 du RGPD. Cette proposition de code de conduite multipartite, finalisé en mai 2022, établit clairement que les plateformes peuvent partager des données personnelles avec des tiers à des fins de recherche d'intérêt public, et ce quelle que soit l'identité du tiers faisant la recherche, **à condition que certaines précautions soient prises dans la protection de ces données**

[15] Le Comité européen pour la protection des données se repose sur un faisceau d'indices (§§ 127-129 des Lignes directrices 8/2020 sur le ciblage des utilisateurs des médias sociaux du 13 avril 2021) pour évaluer si les données ont été « *manifestement rendues publiques par la personne concernée* » (ce qui permet au responsable de traitement de justifier au titre d'exception par la collecte de données sensibles – article 9-2 e) RGPD) :

- le paramétrage par défaut de la plate-forme de médias sociaux (c'est-à-dire si la personne concernée a fait une action pour changer ces paramètres privés par défaut en paramètres publics) ; ou
- la nature de la plate-forme de médias sociaux (c'est-à-dire si cette plate-forme a vocation à connecter des connaissances proches avec la personne concernée ou à nouer des relations intimes (comme les sites de rencontres), ou s'il est destiné à fournir un plus large éventail de relations interpersonnelles, telles que les relations professionnelles, ou le microblogging, le partage de médias, les réseaux sociaux plateformes de partage d'avis en ligne, etc. ; ou
- l'accessibilité de la page où les données sensibles sont publiées (c'est-à-dire si les informations sont accessibles au public ou si, par exemple, la création d'un compte est nécessaire avant d'accéder aux informations) ; ou
- la visibilité de l'information sur le caractère public des contenus publiés par les personnes concernées (exemple : c'est-à-dire s'il y a une bannière continue sur la page ou si le bouton de publication informe la personne concernée que les informations seront rendues publiques…) ; ou
- si la personne concernée a elle-même publié les données sensibles, ou si, à l'inverse, les données ont été publiées par un tiers (exemple : une photo publiée par un ami qui révèle des données sensibles) ou déduites.

Le CEPD note que la présence d'un seul élément peut ne pas toujours suffire à établir que les données ont été « manifestement » rendues publiques par la personne concernée. En pratique, une combinaison de ces ou d'autres éléments peut devoir être prise en compte pour pouvoir démontrer que la personne concernée a manifesté clairement l'intention de rendre les données publiques.

(Vermeulen, 2022 :7). Le projet de code de conduite de l'EDMO vise précisément à définir quelles précautions doivent être prises dans le traitement des données en fonction des risques spécifiques qu'elles présentent. Le projet de code de conduite de l'EDMO s'attèle aussi à définir les responsabilités des plateformes et des chercheurs lors du partage de données. Ce projet sera examiné plus longuement dans la partie 3a) de cette synthèse, qui porte sur le cadre règlementaire.

La CNIL a quant à elle mené une consultation auprès des chercheurs sur leurs modes d'accès aux données à l'égard du RGPD.[16] Ce travail a débouché sur la **publication de ressources visant à aider les chercheurs** à se mettre en conformité : clarifications des bases légales permettant de traiter les données personnelles, procédures à respecter, présentation des outils disponibles pour protéger les données … En particulier, la CNIL a identifié **trois bases légales** permettant le traitement des données personnelles à des fins de recherche : (1) l'exécution d'une mission d'intérêt public (Art. 6(1)(e)); (2) le consentement des personnes concernées (Art. 6(1)(a)) ; et (3) l'intérêt légitime du responsable de traitement (Art. 6(1)(f)). Ces trois bases légales sont également celles utilisées dans le rapport de l'EDMO (2022 : 27) mentionné ci-dessus.

Les plateformes elles-mêmes portent un regard critique sur la situation actuelle, qui les incite à une certaine frilosité en matière de partage de données en raison du manque de visibilité sur le partage des responsabilités légales en cas de fuite de leur données suite au partage avec un chercheur tiers. Les plateformes justifient notamment leur prudence en faisant valoir qu'en cas de fuite, il est impossible d'exercer un contrôle sur l'utilisation potentiellement malveillante qui est faite des données. Les plateformes déclarent également ne pas vouloir s'ingérer dans les accès obtenus par les chercheurs pour ne pas être soupçonnées d'atteinte à l'indépendance de la recherche.

La consultation a fait ressortir **un large consensus, aussi bien pour les chercheurs que pour les plateformes, sur le fait que les pratiques en matière d'accès aux données pourraient être améliorées.**

3. Périmètre de la régulation et construction d'une nouvelle gouvernance

a) Evolution du cadre règlementaire européen

L'accès aux données des très grandes plateformes et moteurs de recherches en ligne verra ses modalités évoluer considérablement suite à la mise en œuvre du **Règlement sur les Services Numériques** et à la mise en place du **Code de bonnes pratiques contre la Désinformation**. Il est donc capital de procéder à une analyse de ces deux textes, afin de préciser leurs positionnements en matière d'accès aux données des plateformes et d'analyser les enjeux liés à leur mise en œuvre. Il est également nécessaire de considérer les efforts entrepris par l'EDMO conjointement à ces deux textes, pour mettre en place un cadre d'accès des chercheurs aux données des plateformes qui sont considérées comme sensibles, notamment parce qu'elles comportent des données personnelles pour lesquelles les utilisateurs pouvaient s'attendre raisonnablement à ce qu'il n'y ait pas de traitement, et dont la fuite pourrait porter atteinte aux droits des individus concernés.

[16] Recherche scientifique (hors santé) : quelle base légale pour un traitement de recherche ? | CNIL

Le RSN et le Code de bonnes pratiques, destinés à fonctionner de manière complémentaire, sont néanmoins de natures différentes. En effet, le RSN introduit un certain nombre d'obligations légales, alors que le Code de bonnes pratiques consiste en une série d'engagements signés volontairement par plusieurs grandes plateformes à l'issue d'un processus de concertation avec des associations, des chercheurs, des annonceurs et la Commission européenne.

De façon remarquable, le RSN vient bouleverser le régime actuel de partage de données, dans la mesure où ce seront les **régulateurs**, et **non plus les plateformes**, qui prendront la décision finale d'accorder l'accès aux données ou non en suivant des critères clairs fixés par le texte. Ce changement de paradigme était réclamé depuis plusieurs années par les chercheurs et même par certains opérateurs, et emporte l'adhésion d'une grande partie des répondants à la consultation. Dans les faits, le succès et la durabilité du RSN reposeront en partie sur le travail mené par les chercheurs sur les risques systémiques générés par les plateformes, travail qui viendra guider la régulation (Husovec, 2022).

Le Règlement sur les Services Numériques

Le RSN vise à garantir la sécurité des utilisateurs européens et à protéger leurs droits fondamentaux en ligne. Pour ce faire, il fixe des obligations de moyens et de transparence aux opérateurs de services numériques. Partant du constat que tous les services ne présentent pas le même degré de risque, le RSN définit les obligations de chaque opérateur de manière proportionnée à leur taille et aux caractéristiques spécifiques de ses services.

Les très grandes plateformes et moteurs de recherche[17] (VLOPSEs – *Very Large Online Platforms and Search Engines*), considérées comme présentant le plus de risques, sont concernées par des obligations spécifiques, y compris en matière de transparence. Innovation majeure du texte, le respect de ces obligations sera supervisé directement par la Commission européenne, en cohérence avec l'influence transfrontalière et systémique des VLOPSEs. Le principe de pays d'établissement subsiste à travers la désignation d'un **Coordinateur pour les Services Numériques** (CSN, *Digital Services Coordinator* ou DSC en anglais) par chaque état-membre. Les CSN seront en charge de centraliser les retours des diverses autorités compétentes au niveau national, ainsi que d'assurer la liaison avec la Commission et les CSN des autres états-membres.

L'accès aux données des plateformes dans le RSN

Le RSN met en place un **régime d'accès aux données différencié entre le public et les chercheurs dits « agréés »**. En pratique, les opérateurs concernés devront mettre (1) un certain nombre de nouvelles données à disposition du public, tandis que (2) d'autres données, notamment certaines considérées comme plus « sensibles », pourront devenir accessibles aux chercheurs dont la demande d'agrément aura été acceptée par les autorités publiques concernées. Le RSN précise également, de façon ponctuelle, les modalités techniques d'accès, encourageant par exemple la mise en place de fonctions de recherche, l'utilisation de formats exploitables par les chercheurs, ou encore l'amélioration par les plateformes de leurs services d'accès aux données tels que les API[18].

[17] Sont considérées comme VLOPSES les plateformes ou moteurs de recherche dépassant le seuil de 45 millions d'utilisateurs actifs mensuels dans l'Union Européenne.
[18] Le RSN semble faire référence aux API lorsqu'il mentionne la notion d'accès « en temps réel », dans la mesure où un tel accès nécessite l'existence d'API.

Les données et informations mises à disposition du public

Les données et informations que les plateformes devront rendre **publiques** portent notamment sur les pratiques de modération, les systèmes de recommandation et les contenus publicitaires diffusés sur leurs services (Encadré 4). Ces informations dont certaines n'étaient jusqu'ici que difficilement accessibles, présentent un intérêt certain en matière de recherche.

Encadré 4
Aperçu des informations qui devront être rendues publiques par les plateformes

Pour toutes les plateformes en ligne

Les intermédiaires devront inclure dans leurs conditions générales des **informations sur les politiques, procédures, mesures et outils utilisés à des fins de modération des contenus**, y compris la prise de décision fondée sur des algorithmes et le réexamen par un être humain, ainsi que sur le règlement intérieur de leur système interne de traitement des réclamations (Art. 14).

Les opérateurs de plateformes devront fournir des rapports précis sur leurs **efforts de modération des contenus** (comprenant un nombre important d'informations et de chiffres) et indiquer leur **nombre d'utilisateurs mensuels.** Ces rapports devront être publiés chaque année par les plateformes, et au moins tous les 6 mois par les VLOPSEs sous forme de fichier lisible par une machine (Art.15 et 24).

Les opérateurs de plateformes utilisant des **systèmes de recommandation** devront établir dans leurs conditions générales « les principaux paramètres utilisés par ces systèmes » ainsi que « les raisons de l'importance relative de ces paramètres » (Art. 27)

Pour les VLOPSEs uniquement

Les opérateurs de VLOPSEs devront publier au moins tous les 6 mois des données additionnelles sur leurs **pratiques de modération** précisant les ressources humaines qu'ils y consacrent pour chacune des langues officielles de l'Union, ainsi que des précisions sur les statistiques tirées de leurs procédures de modération (efficacité, comparaisons par langages modérés, …). Ils devront aussi mettre à disposition du public les **rapports d'audits** auxquels ils ont été soumis, exposant entre autres les résultats de l'évaluation des risques, les mesures d'atténuation mises en place et l'implémentation des recommandations d'audit (Art. 37 et 42).

Les opérateurs de VLOPSEs devront également tenir sur leur interface en ligne un registre contenant un grand nombre d'informations sur les **publicités ayant été diffusées sur leurs services** (financeurs, ciblage, contenu, période, diffusion…), tout en veillant à que ce registre ne contienne pas de données personnelles. Ce registre devra intégrer « un outil de recherche fiable permettant d'effectuer des recherches multicritères et être accessible par le biais d'une API (Art. 39). Le RSN précise que des lignes de conduites relatives à la structure et les fonctionnalités de ces registres publicitaires pourraient être publiées après consultation de chercheurs agréés (Art. 39-3).

Protection des API de recherche et incitation à leur généralisation

Autre évolution notable : le RSN précise que les plateformes devront faciliter l'accès **en temps réel** des chercheurs (pas seulement ceux ayant obtenu l'agrément) **aux données qui sont publiquement accessibles** sur leur interface en ligne dans la mesure où celles-ci contribuent à la lutte contre les risques systémiques (Art. 40-12). Cette disposition spécifique a pour but de formaliser les accès aux données publiques

permis par les API existantes, et *a fortiori* de requérir la mise en place de tels outils par les VLOPSEs qui ne l'auraient pas encore fait.

Le considérant 98, qui complète l'article 40-12, précise également que les opérateurs ne devraient pas « empêcher les chercheurs d'utiliser » les données disponibles publiquement sur leurs interfaces à des fins de compréhension des risques systémiques. Ce considérant pourrait permettre au régulateur de contester les Conditions Générales d'Utilisation ou la tarification d'API considérées comme antagonistes avec la recherche sur les risques systémiques, et pourrait éventuellement fournir une protection aux chercheurs utilisant le *scraping* de données.

Les données à destination de chercheurs agréés

Les **chercheurs agréés** auront quant à eux la possibilité de demander un accès étendu aux données des VLOPSEs par le biais d'une procédure spécifique impliquant directement le régulateur (détaillée ci-dessous). **Ces accès seront circonscrits aux données nécessaires pour des projets de recherche *« contribuant à la détection, au recensement et à la compréhension des risques systémiques*** (voir définition ci-dessous) *dans l'Union […] ainsi qu'à l'évaluation du caractère adéquat, de l'efficacité et des effets des mesures d'atténuation des risques* » pris en vertu du RSN (Art. 40-4). Cette possibilité d'accès étendu placera les chercheurs au cœur de la bonne mise en place du texte. En effet, leurs travaux joueront un rôle-clé en appui du travail des régulateurs, des auditeurs, et des plateformes elles-mêmes.

La notion de risque systémique est centrale dans la **définition du périmètre des données** auxquelles les chercheurs agréés pourront accéder, dans la mesure où cette notion rentrera en ligne de compte lors de l'agrément des projets de recherche et dans la définition des données accessibles. Si le RSN définit dans une certaine mesure les différents types de risques systémiques, un certain nombre de questions subsistent quant à l'opérationnalisation de ce concept et son application aux problématiques de recherche (voir encadré 5).

Encadré 5
La notion de risque systémique dans le RSN (Art. 34)

Quatre types de risques systémiques sont identifiés comme pouvant résulter de la conception, du fonctionnement et de l'usage fait des services numériques[19] :

1) La diffusion de **contenus illicites** (dans le droit des états-membres) ;

2) « *Tout effet négatif réel ou prévisible* » pour l'exercice des **droits fondamentaux**, notamment le droit à la dignité humaine, le droit à la vie privée, à la non-discrimination, la liberté d'expression et d'information, ainsi que la protection des données personnelles, des mineurs et du pluralisme des médias ;

3) « *Tout effet négatif réel ou prévisible sur le* ***discours civique****, les* ***processus électoraux*** *et la* ***sécurité publique*** » ;

4) « *Tout effet négatif réel ou prévisible* » sur la **santé publique et mentale** des individus.

Opérationnalisation du concept de risque systémique

La notion de risque systémique devra être affinée dans le temps, au fur et à mesure de l'accumulation de connaissances sur les services des plateformes en ligne et des décisions du régulateur. Les chercheurs sont donc amenés à jouer un rôle dans l'opérationnalisation du concept de risque systémique au cours du temps et dans l'évolution du périmètre des données éligibles aux accès. Le considérant 97 fournit quelques exemples de données auxquelles les chercheurs agréés pourraient avoir accès, citant notamment le nombre de vues ou les contenus retirés par les opérateurs.

A ce stade, deux approches semblent se distinguer s'agissant des projets de recherche sur les risques systémiques :

- L'étude des **facteurs de risque systémiques consubstantiels aux systèmes des VLOPSEs** (étude sur les biais algorithmiques, les failles dans la modération, des effets indésirables de certains modèles d'affaires et certaines fonctionnalités etc…)
- L'étude de **phénomènes de grande échelle résultant d'actions d'utilisateurs** sur les systèmes des plateformes (terrorisme en ligne, manipulation de processus démocratiques, exploitation de mineurs, haine en ligne, etc…)

Bien que cette distinction conceptuelle soit pertinente, il est important de noter que ces deux approches restent liées, dans la mesure où les actions d'utilisateurs interagissent avec les systèmes des VLOPSEs. Un bon exemple, mentionné à plusieurs reprises dans le RSN, est celui des « opérations coordonnées visant à amplifier l'information » (par exemple dans le considérant 104), qui résultent simultanément d'actions d'utilisateurs et de caractéristiques des systèmes, dans la mesure où les utilisateurs peuvent exploiter les faiblesses de ces systèmes.

[19] Le considérant 79 précise la notion de risque systémique en faisant appel à la notion de *gravité* , *d'irréversibilité*, et la *probabilité* de ces risques. Le considérant 80 mentionne quant à lui l'idée d'une *propagation large et rapide* des conséquences négatives afin de préciser l'appréciation qui sera faite des risques systémiques.

Exemples de sujets de recherche portant sur les risques systémiques

Un premier exemple, déjà identifié par les chercheurs, consiste en l'étude de l'existence, l'ampleur et les conséquences des « **bulles informationnelles** », qui peuvent avoir des effets sur la liberté d'information et sur nos processus démocratiques. A l'avenir, on pourrait par exemple imaginer que le DSA puisse permettre d'étudier **l'influence de la publicité en ligne sur l'alimentation** des plus jeunes, dans la mesure où les risques systémiques couvrent les problématiques de santé publique.

Le statut de chercheur agréé sera délivré selon un processus établi dans l'article 40-8 du RSN, qui comporte **une liste de critères exhaustifs préalables à l'agrément** (détaillés dans l'encadré 6). Ces différents critères s'ajoutent à la nécessité que les projets de recherche portent sur les risques systémiques ou sur l'évaluation des mesures d'atténuation des risques (encadré 5). Cette évaluation des requêtes d'accès aux données, ainsi que les accès qui en résulteront, seront spécifiques à chaque projet de recherche.

Encadré 6
Conditions nécessaires afin d'obtenir le statut de chercheur agréé (Art.40-8)

1) Être affilié à un organisme de recherche ;
La notion d'« organisme de recherche » n'est pas limitée ici aux seules institutions académiques. En effet, elle inclut également les instituts de recherche non-académiques et « tout autre entité ayant pour objectif premier de mener des recherches scientifiques, ou d'exercer des activités éducatives comprenant également des travaux de recherche scientifique », ce qui semble inclure certaines associations de la société civile (Considérant 97 du RSN ; Art. 2 de la Directive européenne sur le droit d'auteur).

2) Être indépendant d'intérêts commerciaux ;

3) Indiquer la source de financement de ses recherches ;

4) Justifier de sa capacité à respecter les exigences spécifiques en matière de sécurité et de confidentialité des données, ainsi qu'à protéger les données personnelles ;

Les chercheurs devront pour cela décrire les mesures techniques et organisationnelles mises en place à cet effet

5) Démontrer que les données concernées et que la période d'accès sont proportionnées au projet de recherche, et que les fins de la recherche sont alignées avec **la lutte contre les risques systémiques** ;

6) S'engager à mettre gratuitement à disposition du public les résultats de la recherche dans un « délai raisonnable » après l'achèvement de celle-ci, sous réserve de respect du RGPD.

Les demandes pourront être déposées par les chercheurs auprès du **CSN de l'état-membre auquel leur organisme de recherche est affilié**, ou bien directement auprès du **CSN d'établissement des VLOPSEs concernées**. Dans le premier cas, le CSN de l'état-membre où les chercheurs sont établis procèdera à une évaluation initiale de la requête d'accès avant de transmettre la demande au CSN d'établissement de la VLOPSE, accompagnée d'un avis sur les suites à y donner. Cet avis devra être pris en compte par le CSN d'établissement des VLOPSEs (Art. 40-9 RSN). Cette possibilité est

pertinente dans la mesure où le CSN de l'état-membre où sont basés les chercheurs postulants à l'agrément est plus susceptible de disposer de certaines données utiles à l'examen de la demande.

Pour leur part, **l'examen des requêtes et la décision finale d'agrément relèveront nécessairement des prérogatives des CSN du pays d'établissement** des VLOPSEs concernées.[20] Si le texte ne mentionne pas explicitement de délais contraignants, il requiert toutefois que l'examen des demandes d'agrément se déroule dans « les meilleurs délais » (Art 40-9). Dans les faits, les CSN d'établissement des VLOPSEs devront donc mettre en place des processus robustes afin de traiter les demandes d'agrément dans des délais acceptables et d'éviter les blocages. Comme indiqué ci-dessus, les CSN des chercheurs auront également la possibilité de jouer un rôle en appui des CSN d'établissement des VLOPSEs afin de faciliter le traitement des demandes.

En cas de décision favorable de leur CSN d'établissement, les VLOPSEs seront tenues de fournir aux chercheurs agréés l'accès aux données. **Les opérateurs auront toutefois la possibilité de proposer des modifications** « *s'ils considèrent ne pas être en mesure de fournir l'accès aux données demandées* ». Deux raisons sont considérées valables à cet égard : (1) les opérateurs concernés n'ont pas accès aux données demandées, (2) fournir l'accès occasionnerait « *d'importantes vulnérabilités pour la sécurité de leur service ou la protection d'informations confidentielles, en particulier des secrets d'affaire* » (Art 40-5). Le CSN concerné aura ensuite 15 jours pour statuer sur le bien-fondé des demandes de modification des opérateurs (Art. 40-6) mais aura de fait le dernier mot pour juger du bien-fondé des demandes d'agrément (voir encadré 7).

Le CSN ayant accordé l'accès garde par la suite **la possibilité de révoquer** celui-ci à tout moment s'il estime que les conditions ayant donné lieu à l'accès ne sont plus remplies (Art. 40-10). Le cas échéant, le chercheur agréé dont l'accès est révoqué a toutefois la possibilité de réagir aux conclusions du CSN.

Encadré 7
Prise en compte des demandes de modification des plateformes concernant les demandes d'accès des chercheurs dans le RSN

Le RSN permet la prise en compte des objections opposées par les opérateurs, notamment celles se basant (1) sur **l'indisponibilité des données**, et (2) sur des arguments de **protection de la confidentialité**, de la **sécurité** et du **secret des affaires** (évoqués dans la partie 2b) de cette synthèse). Dans les faits, la proportionnalité des demandes fera également l'objet d'une attention particulière.

Le CSN d'établissement des VLOPSEs, en tant que décisionnaire final en matière d'agrément, devra trancher quant au bien-fondé des objections des plateformes.[21] Le CSN d'établissement devra s'assurer que les demandes respectant les conditions d'agrément débouchent sur un accès, tout en préservant au mieux les intérêts légitimes de chacun. Ainsi, le considérant 97 précise que « la prise en compte des intérêts commerciaux des fournisseurs ne devrait pas conduire à un refus d'accès aux données nécessaires à l'objectif de recherche spécifique lié à une demande » venant de chercheurs agréés au titre du RSN.

[20] CSN irlandais pour Facebook, Instagram, Twitter, Google, YouTube ; CSN luxembourgeois pour Amazon, Bing
[21] Le CERRE a notamment exploré les considérations d'équilibrage qui devront être prises en compte par le CSN d'établissement dans l'examen des demandes d'agrément (Edelson et al., 2023 : 30)

Dans les faits, **le régulateur aura donc la possibilité de refuser les propositions de modification des opérateurs** et d'imposer l'accès des chercheurs agréés aux données concernées par leurs demandes, notamment en s'appuyant sur **les mesures techniques de sécurisation** des données en réponse aux objections des plateformes. Il sera donc important de suivre les décisions des CSN quant aux demandes de modification des plateformes dans le cadre des demandes d'agrément.

Le RSN précise également que la Commission adoptera un acte délégué afin de préciser les « *conditions techniques dans lesquelles les fournisseurs de très grandes plateformes en ligne ou de très grands moteurs de recherche en ligne partagent des données* » en vertu de l'article 40 et « *les fins auxquelles ces données peuvent être utilisées* ». En pratique, cet acte délégué viendra notamment préciser les modalités techniques permettant un accès conforme avec le RGPD, notamment en matière de **sécurisation** des données (l'exemple des « coffres de données » est notamment donné dans le considérant 97), de **minimisation** des données partagées et **d'anonymisation** de celles-ci. La définition et la mise en place de mesures techniques adaptées à l'accès et au traitement de données dites sensibles pourrait permettre de faire évoluer le périmètre des accès dans le temps.

L'acte délégué précisant l'application de l'article 40 pourrait également venir formaliser le recours à un « ***mécanisme consultatif indépendant*** *à l'appui du partage de données* », qui apporterait une expertise technique additionnelle dans l'examen des demandes d'agrément. Si un tel intermédiaire n'existe pas encore, il semble se rapprocher conceptuellement du tiers de confiance actuellement mis en place par le groupe de travail de l'EDMO sur l'accès des chercheurs aux données des plateformes suite aux engagements pris par les signataires du Code de bonnes pratiques contre la Désinformation (voir encadré 9). Le CSN en charge de l'agrément aurait alors la possibilité de consulter ce tiers. Le recours à un tiers indépendant pourrait notamment faciliter les compromis suite aux demandes de modifications des plateformes dans le cadre des demandes d'agrément.

Cet acte délégué ne pourra être adopté qu'après **consultation préalable du Comité Européen pour les Services Numériques**, qui regroupera les CSN des états-membres une fois qu'ils auront été désignés (au plus tard le 17 février 2024, conformément à l'article 49 du RSN). Il devrait donc entrer en vigueur au printemps 2024.

La Commission européenne a ouvert une consultation publique[22] sur ce projet d'acte délégué au printemps 2023 qui s'est clôturée le 31 mai 2023. L'Arcom y a répondu[23] à travers un document anticipant la publication de la présente synthèse.

Nouveau code de bonnes pratiques contre la désinformation

De façon similaire au RGPD, le RSN encourage la mise en place d'initiatives additionnelles ciblées – dites de ***co-régulation*** –entre les parties prenantes concernées telles la mise en place de **standards** et de **codes de conduite** (Articles 44 à 48 du RSN). Ces codes de conduite pourraient à terme s'articuler avec le RSN, dans la mesure où leur respect serait encouragé (car ils signaleraient une volonté de bien faire), et évalué dans le cadre des mesures de supervision s'appliquant aux VLOPSEs (voir le considérant 104 du RSN et la page 2 du code). L'objectif affiché est que les meilleures pratiques soient construites par les parties prenantes et qu'elles se généralisent rapidement au sein de l'industrie.

[22] Consultation publique sur l'acte délégué sur l'accès aux données par les chercheurs prévu par le DSA
[23] Réponse de l'Arcom à la consultation publique de la Commission européenne

C'est précisément dans cette logique que s'inscrit le **Code de bonnes pratiques contre la désinformation**, signé à l'été 2022 et cité par le considérant 106 du RSN comme une composante importante des efforts d'autorégulation des opérateurs. Ce texte, qui est une version renforcée d'un premier document datant de 2018, consiste en une série d'engagements pris par certains opérateurs de services en ligne dont Meta (Facebook et Instagram), Google (Google Search et YouTube), Microsoft (Microsoft Bing et LinkedIn) et TikTok[24], ainsi que d'autres parties prenantes telles des associations de la société civile, des fédérations d'annonceurs, ou des acteurs spécialisés dans le fact-checking.[25]

Chacun des engagements pris par les signataires du Code de bonnes pratiques est assorti **d'éléments concrets de *reporting* qui devront être communiqués de façon biannuelle** par les signataires. Ces rapports des signataires, dont la première édition a été publiée en février 2023, sont mis à disposition du public via un nouveau portail internet, suivant le principe de guichet unique. Ce portail, appelé « centre de transparence »[26], est lui-même cogéré par les signataires, qui se sont formellement engagés à le tenir à jour conformément à la section 8 du Code. En parallèle, le Code établit un groupe de travail mené par la Commission européenne qui sera chargé de contrôler le respect des obligations, d'harmoniser les pratiques de *reporting*, et d'évaluer l'efficacité des mesures de façon conjointe avec les régulateurs nationaux et européens (section 9 et 10 du Code).

Accès des chercheurs aux données des plateformes signataires du code

Le Code de bonnes pratiques souligne explicitement le rôle-clé des chercheurs dans la lutte contre la désinformation, consacrant une section entière à ce sujet (section 6). Comme le RSN, le Code distingue (1) l'accès des chercheurs aux données **publiques** de (2) l'accès aux **données considérées comme plus sensibles**, qui nécessitera un agrément. Si les deux textes adoptent des approches similaires quant à la qualification des chercheurs, ils diffèrent en revanche dans leur définition du périmètre des projets de recherche éligibles aux accès. Ainsi, quand le RSN fait référence aux recherches portant sur les risques systémiques associés aux plateformes en ligne, le Code de bonnes pratiques restreint quant à lui son champ d'application aux seules recherches portant sur la désinformation, ce qui pourrait avoir des implications sur l'attribution des accès (voir encadré 8).

[24] Twitter, qui faisait initialement partie des signataires, s'est retiré du Code en mai 2023.
[25] Voir la liste complète des signataires : Signatories of the 2022 Strengthened Code of Practice on Disinformation | Shaping Europe's digital future (europa.eu)
[26] Accessible à l'adresse suivante : Home - Transparency Centre (disinfocode.eu)

Encadré 8
Périmètre des accès permis par le RSN et le Code

Eligibilité des chercheurs

Comme le RSN, le Code de bonnes pratiques ne restreint pas la qualification de « chercheur » aux seuls universitaires, et inclut également des chercheurs venus de la société civile, s'ils satisfont les exigences applicables en matière d'indépendance financière, de définition de leur projet de recherche, et de respect des bonnes pratiques éthiques et méthodologiques (Code de bonnes pratiques, 2022 : 26-27)

Risque systémique (RSN) & Désinformation (Code)

Le champ d'action du Code de bonnes pratiques est en apparence plus restreint que celui du RSN, dans la mesure où la désinformation - ne constitue qu'une facette du concept de risque systémique cité dans le RSN. Inversement, il semble que le Code de bonnes pratiques couvre les problématiques de désinformation sans mentionner leur caractère systémique. Il est donc difficile de prédire avec certitude le degré de variation qui existera entre les deux textes en termes d'éligibilité des projets de recherche.

Les signataires du Code s'engagent à partager publiquement (i.e. pas seulement avec les chercheurs) les données « non-personnelles » et les données « anonymisées, agrégées ou manifestement rendues publiques » servant à la recherche contre la désinformation, citant en exemple les données d'engagement et les impressions (engagement 26.1). Les signataires s'engagent aussi à fournir, cette fois **aux seuls chercheurs s'intéressant à la désinformation**, un « accès en temps réel, ou quasi-temps réel, et lisible par une machine, aux données non-personnelles et aux données publiques anonymisées, agrégées ou manifestement rendues publiques », telles que les données associées aux comptes de personnalités publiques ou les comptes gouvernementaux (engagement 26.2). Ces accès en temps réel, qui ne sont pas toujours possibles actuellement, seraient soumis à des processus de demande « qui ne seraient pas excessifs », probablement similaires aux processus existants d'accès aux API destinées aux chercheurs.

Ces engagements font écho à **l'article 40-12 du RSN** mentionné précédemment, qui vise à protéger les accès aux données publiques *via* les API existantes, et *a fortiori* à requérir la mise en place de tels outils par les signataires qui ne l'auraient pas encore fait.

En parallèle, d'autres engagements du Code pourraient aussi mener à l'ouverture de **nouvelles données** au public et aux chercheurs. C'est notamment le cas des engagements 10 & 11 (mise en place de **registre des publicités à caractère politiques** – faisant écho à l'article 39 du RSN), de l'engagement 18.3 (investissement dans la recherche sur le design sûr) et de l'engagement 31 (mise en place d'un registre du travail des *fact-checkers*).

Pour les données considérées comme plus sensibles, le Code recourra à une **procédure d'agrément des chercheurs qui est pour l'instant différente de celle du RSN**, dans la mesure où elle devrait être supervisée par un tiers-parti indépendant utilisant la méthodologie présentée par le projet de code de conduite de l'EDMO (voir encadré 9). La principale différence entre ces deux mécanismes d'accès aux données considérées comme sensibles réside donc dans leur **gouvernance** : dans le RSN, l'accès se fera en *réaction* à une décision d'agrément venue du régulateur, tandis que les accès fournis dans le cadre du Code de bonnes pratiques devraient se faire de manière bilatérale et

proactive entre plateformes signataires et chercheurs, avec une **médiation du tiers indépendant en cas de difficulté** pour trouver un arrangement (Vermeulen, 2022 :5 ; EDMO, 2022 : 55).

Ces mécanismes s'articuleraient idéalement de manière efficace à terme : un certain nombre d'accès seraient ainsi accordés par le tiers de confiance (et donc sans intervention directe du régulateur), permettant de **répartir la charge de travail entre le tiers de confiance et le régulateur** (en particulier le CSN d'établissement des VLOPSEs). A terme, l'acte délégué qui viendra compléter l'article 40 du RSN pourrait donner un rôle consultatif au futur tiers indépendant de l'EDMO dans l'examen des demandes d'agrément au sens de l'article 40 du RSN.

Encadré 9
Code de conduite de l'EDMO et mise en place d'un tiers indépendant

Les opérateurs de services signataires de l'engagement 27 du Code de bonnes pratiques (parmi lesquels figurent Meta, TikTok, Microsoft et Google) s'engagent à « fournir aux chercheurs agréés l'accès aux données nécessaires pour entreprendre des recherches sur la désinformation », et à « développer, financer et coopérer avec un tiers de confiance indépendant » qui « pourra agréer les chercheurs et les propositions de recherche ».

Cet engagement cite à plusieurs reprises l'initiative du **groupe de travail de l'EDMO sur l'accès des chercheurs** aux données, expliquant que le projet de code de conduite de ce groupe de travail pourrait guider la constitution et l'action subséquente de ce tiers indépendant. Ce groupe de travail de l'EDMO, qui incluait notamment Google, Twitter et Meta, est d'ailleurs mentionné dans plusieurs des rapports de transparence publiés par les signataires en février 2023, illustrant son implication dans la création du tiers indépendant mentionné dans le code.

Pour rappel, le projet de code de conduite de l'EDMO vise à clarifier les mesures permettant de partager des données personnelles à des fins de recherche de façon conforme au RGPD et à définir les responsabilités de chaque acteur impliqué dans le partage et le traitement des données. Dans cette optique, le code de conduite de l'EDMO établit un cadre permettant (1) d'évaluer le niveau de risque associé à l'accès et à l'analyse de catégories spécifiques de données et (2) de définir des processus techniques clairs et adaptés afin d'atténuer ces risques et donc de permettre l'accès aux données concernées aux chercheurs (Vermeulen, 2022 : 7).

Le code de conduite de l'EDMO appelle explicitement à la création d'un tel tiers, dans la mesure où il permettrait la reconnaissance officielle de ce code de conduite selon le processus défini dans **l'article 40 du RGPD**, et où il permettrait de superviser l'agrément des chercheurs de manière efficace, indépendante et conforme au code de conduite (EDMO, 2022 : 12). Un tel tiers de confiance disposerait notamment d'une expertise spécifique dans l'examen méthodologique des projets de recherche et dans la maitrise des diverses mesures techniques permettant de sécuriser les données personnelles. Le document de l'EDMO précise que ce tiers de confiance pourrait aussi jouer à terme un rôle consultatif dans le **processus d'agrément des chercheurs du RSN**, et prévu par son l'**article 40-13** mentionné précédemment.

Le groupe de travail de l'EDMO liste également d'autres fonctions que ce tiers pourrait remplir, dont certaines font d'ailleurs écho à des engagements pris dans le cadre du code de bonnes pratiques (EDMO, 2022 :13) :

- Supervision des **documents d'information** publiés par les plateformes à destination des chercheurs ;

- Information des chercheurs quant à la **disponibilité des données** ;

- Mise en place d'un **portail répertoriant les différents projets de recherche** faisant appel à des accès résultant du code de conduite, et informant les individus dont les données personnelles pourraient être concernées ;

- Facilitation de l'identification des **besoins en nouvelles données** et de la mise à disposition de ces données.

Si ce tiers indépendant reste pour l'instant à l'état de projet, le groupe de travail de l'EDMO pourrait notamment travailler à la mise en œuvre d'un **projet pilote d'agrément** permettant de tester la faisabilité de son code de conduite. Un tel projet pilote consisterait vraisemblablement à partager un nombre limité de données personnelles à des chercheurs partenaires. La participation à de tels projet pilotes est d'ailleurs mentionnée dans le Code de bonnes pratiques (engagement 27.4), qui cite le « contenu ayant été supprimé des services signataires » comme un exemple de données pouvant être partagées à l'occasion d'un projet pilote.

En sus des engagements concernant la mise à disposition de données, les plateformes signataires se sont également engagées à « soutenir la recherche de bonne foi s'intéressant à la désinformation impliquant leurs services » (engagement 28). Cet engagement implique notamment la mise en place par les signataires de ressources humaines suffisantes permettant de « faciliter la recherche et maintenir un dialogue ouvert avec les chercheurs » pour « suivre les types de données susceptibles d'être demandées par ces chercheurs ». Les signataires s'engagent également à « être transparents au sujet des types de données qu'ils rendent disponibles », à financer la recherche indépendante sur la désinformation en coopérant avec la communauté de recherche européenne et l'EDMO en allouant les fonds de manière transparente et basée sur le mérite scientifique.

Les signataires s'engagent aussi « à ne pas interdire ou décourager la recherche de bonne foi et d'intérêt public véritable et manifeste sur la désinformation concernant leurs services » et « à ne pas prendre de mesures hostiles à l'encontre des utilisateurs chercheurs ou des comptes » prenant part à de telles recherches (engagement 28.3). Ce dernier engagement, pourrait notamment concerner les chercheurs adoptant des **méthodes de recherche basées sur des accès adversariels** au premier rang desquelles figure le *scraping* ou le *crowdsourcing* de données, qui aboutissent parfois à des sanctions (notamment des fermetures de compte) de la part des plateformes (voir la section 2 de cette synthèse). Cette mesure doit d'ailleurs faire l'objet de consultations annuelles organisées en collaboration avec l'EDMO, qui interrogeraient les chercheurs concernés par des mesures de rétorsion de la part des plateformes.

Un dernier engagement pertinent, pris cette fois par des **associations de la société civile et des acteurs spécialisés dans le fact-checking**, consiste en l'adoption systématique de **méthodologies transparentes** et respectant des **standards éthiques**. Cet engagement inclut également le partage de bases de données, et des conclusions de recherche avec les « publics concernés », notamment les régulateurs mais aussi l'EDMO, l'ERGA et les autres signataires du Code, dans le but d'informer les future politiques régulatoires et afin de généraliser les bonnes pratiques. Cet engagement mentionne également la publication, dans la mesure du possible, des conclusions sur le « centre de transparence » en ligne.

En conclusion de cette section portant sur l'évolution du cadre règlementaire, il est important de noter que la mise en place des mécanismes d'accès des chercheurs aux données des plateformes est toujours en cours. Ainsi, les prochains mois seront

déterminants, notamment s'agissant (1) de la publication de l'acte délégué complétant l'article 40 du RSN, (2) de la traduction des engagements des signataires du Code de bonnes pratiques en mesure concrètes et (3) de la formation d'un tiers indépendant d'agrément des chercheurs mettant en œuvre le code de conduite de l'EDMO.

b) Construction d'une nouvelle gouvernance

La consultation, conduite en parallèle de la refonte du cadre de la régulation des plateformes en ligne au niveau européen, a été l'occasion pour l'Arcom de questionner les acteurs sur **la gouvernance qu'ils souhaiteraient voir émerger en matière d'accès aux données**. Cette partie est donc structurée autour de points de consensus qui viendront éclairer une série de propositions de l'Arcom, détaillées dans la partie 4 de cette synthèse.

Sous-optimalité de la situation actuelle

Le premier point de consensus concerne **la sous-optimalité de la situation actuelle**. En effet, les plateformes décident jusqu'ici unilatéralement d'accorder ou de refuser les accès des chercheurs à leurs données. Tous les répondants, plateformes comme chercheurs, s'accordent ainsi pour dire que les pratiques actuelles pourraient être améliorées.

Les chercheurs répondants déplorent les obstacles auxquels ils font actuellement face : disparité des accès entre plateformes, politiques d'accès restrictives, manque de transparence et de certitudes quant aux accès et à leur continuité, outils limités et peu harmonisés, manque de dialogue avec les plateformes… Les chercheurs expliquent que ces obstacles affectent négativement la conception et la mise en œuvre de leurs projets**, retardant ainsi la documentation et la compréhension de certains phénomènes**.

Les plateformes, quant à elles, déplorent notamment **l'absence d'un cadre légal clair en matière de partage de données**, qui les pousse à une certaine frilosité vis-à-vis des demandes d'accès venant des chercheurs. Elles déclarent être prêtes à ouvrir leurs données à condition que la protection des données personnelles des utilisateurs, la sécurité de leurs services et de leurs secrets d'affaires soient prises en compte. Les plateformes avancent également les ressources importantes nécessaires à la mise en place de programmes d'accès et à la bonne mise à disposition de certains nouveaux types de données. De plus, elles soulignent leurs efforts lors des dernières années, avec la mise en place de plusieurs programmes destinés aux chercheurs par YouTube, Facebook et Twitter par exemple.

Ce constat a notamment influencé les modalités de mise en œuvre du RSN, du Code de bonnes pratiques contre la désinformation et du projet de code de conduite de l'EDMO, qui visent à clarifier les conditions de l'ouverture des données des plateformes et à créer une alternative crédible aux modes d'accès discrétionnaires sur décision des opérateurs.

Pertinence d'un accès aux données basé sur le risque

Le second point de consensus concerne la pertinence de l'approche consistant à **adapter les modalités d'accès sur la base du risque présenté par la fuite des données concernées**. Selon cette logique, l'accès à des données dites sensibles devrait se faire selon des modalités permettant davantage de sécurité. Inversement, les **données non-personnelles (ou agrégées)** devraient pour leur part être **ouvertes au plus grand nombre**, en adoptant par exemple les principes de l'*open data,* qui permet la libre mise en place d'initiatives par la société civile. Comme précisé dans la partie 2)b), le niveau

de sensibilité de certaines catégories de données continue néanmoins à susciter le débat, ce qui pourra avoir une influence sur l'évolution des accès dans le temps.

On peut là aussi souligner que l'évolution du cadre règlementaire répond en partie à cette demande, dans la mesure où le RSN et le Code de bonnes pratiques détaillent tous les deux les catégories de données devant être rendues publiques (par exemple, sur les publicités, la modération, les principaux paramètres des systèmes de recommandation), tandis que certains autres types de données, plus sensibles, ne deviendront accessibles qu'aux chercheurs agréés, sur demande de ces derniers. Les deux textes ne fournissent en revanche que peu de détails sur les modalités techniques permettant le partage de données considérées comme sensibles (le RSN mentionne seulement les « coffres de données »), se contentant de souligner la nécessité d'une montée en compétence techniques pour assurer le partage.

Comme évoqué dans la partie précédente, l'acte délégué complétant l'article 40 du RSN et le travail actuel de mise en place d'un tiers indépendant par le groupe de travail de l'EDMO devraient venir apporter des précisions sur ces modalités techniques en matière de **protection,** de **minimisation** et **d'anonymisation** des données partagées avec les chercheurs. Il est néanmoins clair que ce travail devra se poursuivre dans le temps, notamment au fur et à mesure de l'évolution de l'état de l'art en la matière. Certaines institutions de recherche répondantes ont ainsi proposé la mise en place de groupes de travail dédiés au sujet de la protection des données, et se sont déclarées ouvertes à une co-construction des processus techniques d'accès avec les plateformes elles-mêmes. Elles mentionnent par exemple l'idée d'une plateforme sécurisée de partage de données sensibles, ou se réfèrent à des infrastructures de recherche existantes permettant de travailler sur des données sensibles telles le Centre d'Accès Sécurisé aux Données ou l'infrastructure Huma-Num.

En parallèle du partage de données lui-même, la consultation fait également ressortir la question de la **publication et la réutilisation des données de recherche**, nécessaires à la transparence et la réplicabilité des expériences menées par les chercheurs, notamment lors du processus de revue par les pairs. Là aussi, des solutions devront être mise en place à partir d'une concertation entre les acteurs concernés afin de concilier le besoin de transparence et la protection des données concernées.

Le besoin de mécanismes clairs, efficaces et transparents

Montée en puissance des différents mécanismes d'accès

Les chercheurs formulent la demande de processus **standardisés, transparents et clairs,** dont la charge administrative serait limitée**.** Certains soulèvent par exemple la possibilité de continuer à traiter directement avec les plateformes, ne faisant ainsi appel aux mécanismes externes qu'en cas de désaccord ou bien en cas de demande d'accès à des données particulièrement sensibles. L'idée d'une procédure accélérée dans le cas de protocoles de recherche urgents est également évoquée, tandis que certains répondants souhaitent la mise en place de moyens permettant d'assurer une revue périodique du cadre général d'accès aux données afin d'identifier des axes d'amélioration.

Certains échanges ayant eu lieu depuis la publication du RSN et du Code de bonnes pratiques font ressortir l'importance **d'éviter la multiplication de processus parallèles.** Il sera donc important de suivre l'intégration entre le mécanisme d'agrément permis par le RSN et l'agrément des chercheurs mis en œuvre par le tiers indépendant conformément au Code de bonnes pratiques contre la désinformation (voir encadrés 8 et 9 notamment).

Les chercheurs insistent sur l'importance d'un investissement suffisant des plateformes en termes de **moyens humains et matériels** pour fournir une documentation à jour sur les accès à leurs données et la description des évolutions des fonctionnalités de leurs services. Une telle documentation pourrait par exemple inclure les différents types de données disponibles, leurs conditions d'accès et des cas d'usage, regroupés au sein de catalogues de code (« *codebooks* ») publics. Les chercheurs sont demandeurs d'une fonction de support, et notamment la mise en place d'un point de contact identifié au sein des plateformes.

Mise en place d'un tiers indépendant

La consultation sondait explicitement les répondants au sujet d'un possible recours à un **tiers indépendant** en charge d'examiner les demandes d'accès. Si cette solution semble emporter l'assentiment des répondants, il faut toutefois noter quelques désaccords entre plateformes et chercheurs **sur le fonctionnement effectif d'un tel tiers**. Les chercheurs en particulier, redoutent la capture du tiers par les plateformes, et questionnent parfois l'opacité des critères de composition des groupes de travail existants. Ainsi, si les chercheurs acceptent que les plateformes soient impliquées dans la formation du tiers et la définition des protocoles d'agrément, ils sont majoritairement opposés à leur inclusion parmi les décideurs qui évalueront les demandes, mentionnant un risque pour l'indépendance du processus d'évaluation. Les chercheurs craignent également que les plateformes aient des prérogatives de relecture des projets avant leur publication, qui pourraient mener à des pressions. Certains mentionnent un possible droit de retour pour les plateformes, mais davantage à titre d'information, sans leur donner la capacité de bloquer la publication d'un travail de recherche ; un positionnement que l'on retrouve dans le RSN. Une majorité de répondants insiste sur le besoin d'une séparation stricte entre pouvoir décisionnel sur les accès et financement de la recherche.

Enfin, les chercheurs répondants divergent quant au niveau optimal sur lequel devrait se placer un tel tiers : certains proposent **l'échelon national,** qui permettrait une meilleure analyse des spécificités contextuelles, quand d'autres avancent que le niveau **européen serait plus approprié notamment** en matière d'harmonisation des protocoles. Certains chercheurs évoquent la possibilité pour le régulateur d'infliger des sanctions aux plateformes ne se conformant pas aux décisions du tiers de confiance.
Si elles soutiennent l'idée de confier l'agrément à un tiers, les plateformes, quant à elles, font valoir leur droit à protéger leurs données si elles estiment que leur sécurité ou leurs intérêts sont menacés. Elles mettent également en avant les travaux de co-construction avec les chercheurs dans le cadre du groupe de travail de l'EDMO.

Tous les répondants s'accordent en revanche sur la nécessité pour le tiers de rassembler des expertises **techniques**, **méthodologiques** et **juridiques**, afin que l'examen des demandes soit de la meilleure qualité et cohérence possible, que la supervision du respect des obligations soit assurée et que les normes techniques d'accès définies par le tiers soient optimales.

Capitaliser sur les expertises existantes pour contribuer à la vitalité de l'écosystème de recherche

La consultation a permis aux répondants de mettre en avant les **expertises existantes** en matière de partage et de traitement de données, sur lesquelles il sera crucial de s'appuyer afin de permettre la bonne mise en œuvre des nouvelles modalités d'accès aux données, et de faciliter leur appropriation par le plus grand nombre de chercheurs possible.

L'application du RGPD a déjà permis aux institutions de recherche de monter en compétence sur le sujet de la protection des données personnelles depuis 2018.

Nombre d'institutions disposent déjà de **délégués à la protection des données** (DPO en anglais, pour *Data Protection Officers*), formés au respect du RGPD et assurant un rôle de support dans la création ainsi que la mise en place de protocoles de traitement de données. La **CNIL**, en tant qu'autorité en charge de la protection des données, s'est pleinement investie afin d'accompagner les acteurs, notamment par le biais de fiches ressources.

La France et l'Europe disposent de **laboratoires de recherche** produisant des travaux à la pointe de l'état de l'art en matière d'études des plateformes, et dont les contributions à cette consultation ont été d'une grande richesse. Ces institutions jouent déjà un rôle dans la **transmission de savoir et de techniques**, notamment *via* la mise à disposition d'outils visant à faire baisser les coûts d'entrée pour les chercheurs venant d'autres institutions moins habituées à travailler sur le sujet des plateformes, mais également dans la formation d'étudiants.[27] Il sera important de continuer sur cette voie, en facilitant l'échange entre les chercheurs, et notamment en favorisant **l'interdisciplinarité**. La consultation fait également ressortir le besoin d'un dialogue au sein de la communauté de recherche autour de la **hiérarchisation** des besoins en nouvelles données, ainsi qu'en matière de mise à jour et d'harmonisation des outils existants.

Plus généralement, le fait de disposer d'un réseau important de chercheurs qualifiés permettra à ces derniers d'améliorer la robustesse des **processus de revue méthodologique et scientifique par les pairs**, aussi bien lors de la construction des projets de recherche qu'avant leur publication. La revue par les pairs et l'échange entre les chercheurs seront importants pour assurer la **qualité** et **l'indépendance** des travaux.

Les échanges menés en parallèle de la consultation font également ressortir un besoin pour que la **société civile**, au premier rang de laquelle figurent les signaleurs de confiance, puisse jouer un rôle en appui des chercheurs dans la définition des projets. La société civile peut en effet permettre aux chercheurs **d'identifier plus rapidement les problématiques émergentes** nécessitant un éclairage scientifique.

Enfin, il sera capital pour le régulateur d'échanger encore davantage avec le monde de la recherche, afin de créer un **cercle vertueux** permettant aux chercheurs de tirer pleinement parti des nouvelles possibilités d'accès aux données des plateformes, et de produire des travaux de qualité sur lesquels le régulateur pourra s'appuyer.

4. Propositions

La dernière partie de cette synthèse vise à éclairer le rôle que pourrait jouer l'Arcom en **soutien des chercheurs** afin que ceux-ci puissent apporter leurs connaissances au débat public, mais aussi à comprendre comment l'Arcom pourrait **s'appuyer sur les travaux** de recherche pour améliorer sa régulation et nourrir les échanges au niveau européen.

Cette partie est structurée autour de **propositions**, qui font suite aux retours d'expérience des chercheurs et des plateformes, et aux échanges que l'Arcom a eu avec les parties prenantes.

Ces propositions sont structurées autour de deux axes :

[27] Voir par exemple les nombreux outils mis à disposition par le Medialab de Sciences Po. Outils | médialab Sciences Po

(1) Contribution de l'Arcom à la **vitalité de l'écosystème de recherche** en France (partie a).

(2) Contribution de l'Arcom à l'efficacité des mécanismes d'accès aux données mis en place par le **RSN** et le **Code de bonnes pratiques contre la Désinformation** (partie b) ;

Ces deux volets sont en pratique très liés, dans la mesure où ils se renforcent mutuellement. L'objectif de ces propositions est de positionner l'Arcom en tant que partenaire identifié aussi bien pour les autorités compétentes nationales que pour ses partenaires européens, mais aussi de construire une relation de confiance avec le monde de la recherche et les opérateurs de plateformes.

Encadré 10
Indépendance des chercheurs et financement de la recherche

Les actions proposées par l'Arcom respecteront scrupuleusement **l'indépendance des chercheurs**, qui fait leur force. Les initiatives touchant à la structuration de l'écosystème de recherche viendront donc avant tout des chercheurs eux-mêmes. L'Arcom apportera son soutien lorsque celui-ci sera nécessaire et répondra à une demande des chercheurs.

L'accroissement de la quantité et de la qualité des travaux de recherche portant sur les plateformes en ligne passeront également par un développement de **financements adaptés**, question qui ne fait toutefois pas partie du champ des missions de l'Arcom. Il sera important de capitaliser sur les mécanismes de financement existants et de les adapter au défi représenté par les plateformes en ligne. L'Arcom mettra en place un suivi des initiatives en la matière et des difficultés éventuellement rencontrées par les chercheurs tout en restant attentive à ne pas s'ingérer dans l'allocation des financements et dans l'évaluation scientifique des projets de recherche.

a) Contribuer à la vitalité de l'écosystème de recherche français

a.1) Maximiser les synergies au sein de l'écosystème

La France dispose de chercheurs (aussi bien académiques que non-académiques) ayant une grande expérience dans l'analyse des plateformes et la compréhension des phénomènes sociaux qu'elles génèrent. Il est important de pouvoir capitaliser sur les expertises existantes afin d'étendre le réseau de chercheurs et membres de la société civile capable de contribuer à notre connaissance collective des plateformes. Les propositions détaillées dans cette section visent ainsi à faire **baisser le cout d'entrée de la recherche** portant sur les plateformes en ligne et à permettre d'accélérer la **montée en compétence** de nouveaux chercheurs.

- *Faciliter le partage d'expérience entre chercheurs*

L'Arcom se propose d'agir comme facilitateur de la mise en relation et de la circulation de l'information entre les chercheurs travaillant sur les données des opérateurs de plateformes. L'Arcom appelle à la mise en place d'un **registre sécurisé des accès** obtenus par les chercheurs, rempli sur une base volontaire par ces derniers. Un tel dispositif permettrait de centraliser les informations entre les accès obtenus suite à un agrément au sens du RSN, mais aussi suite à des accords bilatéraux entre chercheurs et plateformes, qui seront facilités dans le cadre du Code de bonnes pratiques contre la Désinformation et grâce au code de conduite de l'EDMO. Un tel dispositif faciliterait la

mise en contact et le partage d'expérience entre les chercheurs travaillant sur les mêmes plateformes et des jeux de données similaires, mais également l'élaboration de demandes d'agrément par les chercheurs. Des chercheurs n'ayant pas encore d'expérience avec l'utilisation des données des plateformes (notamment ceux venant d'autres disciplines) pourraient également utiliser ces ressources pour nouer des relations et monter plus rapidement en compétence. L'Arcom souhaite également mettre en avant les divers ateliers méthodologiques déjà mis en place au sein de monde académique, qui permettent à de nombreux chercheurs de se former à l'étude des données des plateformes[28].

➢ *Optimiser la réutilisabilité des jeux de données*

En parallèle, il est important d'optimiser la **réutilisabilité** des jeux de données, dans la mesure où cela améliorerait la fluidité du travail des chercheurs et des plateformes. S'agissant des premières demandes d'agrément au sens du RSN, il serait par exemple opportun de prioriser les demandes aboutissant à la création de jeux de données **les plus demandés** au sein de la communauté de recherche et offrant un **fort potentiel de réutilisation** par les chercheurs européens. Et ce afin de permettre au mécanisme de bénéficier au plus grand nombre de chercheurs le plus rapidement possible.

Les plateformes doivent faire preuve de transparence sur les jeux de données dont elles disposent ou qu'elles ont déjà produits afin de faciliter les demandes des chercheurs. L'Arcom encouragera en parallèle la mise en place de catalogues de code (« *codebooks* ») par les plateformes. Ces documents, qui comporteraient tous les types de données disponibles à un instant *t,* permettraient aux chercheurs d'avoir une meilleure visibilité au moment de formuler leurs demandes d'accès. A terme, ces *codebooks* pourraient être agrégés afin d'offrir un guichet unique aux chercheurs souhaitant réfléchir à la faisabilité de certains protocoles de recherche. L'existence de tels *codebooks* inter-plateformes permettrait également de mettre en avant le manque de disponibilité de certaines données, ainsi que de suivre et comparer les efforts des opérateurs.

➢ *Valoriser l'implication et la montée en compétence de jeunes chercheurs*

L'Arcom souhaite valoriser l'implication **d'étudiants et de jeunes chercheurs** dans l'étude des données issues des plateformes (notamment en échangeant avec les instances compétentes sur la mise en place de filières dédiées par les universités). L'Arcom souhaite également mettre en avant les ressources pédagogiques facilitant l'étude des plateformes (MOOCs, tutoriels…), et soutenir le partage interdisciplinaire entre les sciences de l'informatique et les sciences sociales.

➢ *Valoriser le déploiement d'outils computationnels publics*

De nombreux **outils en accès public** permettent d'exploiter les données des plateformes.[29] Ces outils peuvent faciliter considérablement les travaux de recherche en élargissant l'accès aux individus ne disposant pas de formations en sciences des données et d'attirer des profils plus divers (psychologie, santé, etc.). La généralisation de tels outils est un moyen efficace de faire baisser les coûts d'entrée pour les chercheurs. L'Arcom se propose donc de valoriser et d'appuyer leur déploiement en associant les laboratoires de recherche spécialisés et le PEReN. De tels outils pourraient par exemple faciliter le **traitement et l'analyse d'images et de vidéos** par les

[28] Les ateliers « Metat » organisés par Sciences Po sont un bon exemple.
[29] En France, le Medialab de Sciences Po fait office de figure de proue dans le déploiement et la popularisation de tels outils.

chercheurs, dans la mesure où ceux-ci constituent une part importante des contenus circulant sur les plateformes.

➢ *Soutenir les approches pluridisciplinaires et transnationales*

L'Arcom souhaite encourager les approches **pluridisciplinaires** et **transnationales**, par exemple en facilitant la collaboration avec des chercheurs, associations et régulateurs d'autres pays européens. Il sera pour cela important de développer des échanges réguliers avec les autres CSN, et d'encourager la création de jeux de données permettant de produire de la connaissance sur des phénomènes affectant plusieurs états membres de l'Union.

➢ *Soutenir le déploiement de solutions techniques sécurisées*

L'Arcom se propose d'accompagner le déploiement de solutions techniques sécurisées adaptées au traitement de données dites sensibles à des fins de recherche. La mise en place d'**infrastructures sécurisées et mutualisées** adaptées au traitement de ces données permettra de faire évoluer les accès des chercheurs dans le temps, et de faire baisser le coût d'entrée pour les chercheurs. Il sera pour cela opportun de s'appuyer sur les bonnes pratiques mises en place par le Centre d'Accès Sécurisé aux Données[30] ou l'infrastructure de recherche Huma-Num. Les solutions mises en place par ces acteurs pourraient par exemple être servir d'inspiration à des infrastructures adaptées à l'étude des données des plateformes. La mise en place de tels dispositifs pourrait bénéficier d'une collaboration entre la CNIL, les DPO des universités, les laboratoires de recherche et les plateformes elles-mêmes. Une telle réflexion, si elle veut se développer à l'échelle, doit être portée en coordination avec le niveau européen.

L'Arcom appelle à la mise en place d'un **groupe de travail** dédié à l'identification et au suivi des mesures techniques les plus avancées permettant le partage, l'archivage et la publication des données et des résultats de recherche. Ce groupe de travail, qui réunirait régulateurs (CSN, autorités de protection des données…) et parties prenantes (spécialistes techniques, plateformes, chercheurs, …), aurait pour fonction de suivre l'évolution de l'état de l'art en matière d'anonymisation, de minimisation et de protection des données[31], permettant la mise en œuvre des projets et leur revue par les pairs dans un cadre sécurisé. Une telle connaissance sera importante afin de pouvoir contribuer à l'évaluation des demandes d'agrément.

➢ *Favoriser une logique d'open data s'agissant des données les moins sensibles*

Les répondants ont largement souligné l'importance d'un accès public aux données jugées non-sensibles. Dans ce cadre, l'Arcom souhaite soutenir les démarches d'***open data*** s'agissant aussi bien des données des plateformes que celles des chercheurs (par exemple certaines bases de données préalablement nettoyées, harmonisées et/ou recombinées). L'*open data* permet à la société de s'approprier les questions relatives à la supervision des plateformes. Il est important de ne pas encourager la multiplication des guichets en matière d'*open data* afin de garantir une meilleure visibilité pour les

[30] Le Centre d'accès sécurisé aux données permet à des chercheurs français et européens de travailler sur des données sensibles et très détaillées telles des données administratives ou de santé. La technologie utilisée par le CASD permet une authentification forte des chercheurs, tout en confinant les données et en permettant la traçabilité des accès, garantissant ainsi un très haut niveau de sécurité.
Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique – Courrier des statistiques N3 - 2019 | Insee
[31] Par exemple à la confidentialité différentielle, aux coffres de données et aux API sécurisées.

chercheurs. L'Arcom appelle donc à capitaliser sur les initiatives et infrastructures existantes.

De par sa position d'intermédiaire entre les opérateurs, les chercheurs et le public, l'Arcom pourra pour sa part participer à la **mise en avant des données qui devront être rendues publiques par les plateformes dans le cadre du RSN**.

a.2) Réguler en s'appuyant sur les travaux de recherche

Le succès du RSN dépend de l'implication de toutes les parties prenantes.

L'écosystème de recherche et la société civile peuvent **augmenter la capacité du régulateur** à détecter des problématiques émergentes, à comprendre les besoins en nouvelles données, et à analyser efficacement l'évolution des politiques et des systèmes déployés par les plateformes. Le régulateur, quant à lui, doit pouvoir partager ses conclusions avec les parties prenantes et rendre compte de ses interactions avec les opérateurs, afin de contribuer à la **construction d'une relation de confiance** entre le public et ces dernières. Le régulateur doit également pouvoir mettre en avant auprès des parties prenantes certains sujets méritant un éclairage[32], et rester au fait de l'état de l'art en matière de recherche.
Ce travail aux côtés de l'écosystème de recherche permettra au régulateur de faire remonter les connaissances acquises et les préoccupations des chercheurs auprès des partenaires européens.

➢ *Contribuer à la visibilité des chercheurs et de leurs travaux dans le débat public*

Les connaissances produites par les chercheurs permettent d'éclairer le débat public autour des plateformes en ligne.
L'Arcom se propose de faciliter la mise en place d'un **annuaire public des chercheurs** travaillant sur les données des plateformes, ainsi que de mettre en avant les publications mobilisant les données des plateformes. Ces ressources pourront servir de référence au grand public, aux décideurs et aux médias pour consulter les travaux existants et identifier les chercheurs travaillant sur ces thématiques.

➢ *Identifier les problématiques de recherche émergentes*

Le régulateur doit pouvoir s'appuyer sur les parties prenantes afin d'identifier et de comprendre les **problématiques émergentes** sur les plateformes en ligne. Ce travail permettra de contribuer à la définition de la notion de **risque systémique**.

Le régulateur doit contribuer à mettre en place des formats de dialogue entre la **société civile** (en particulier les signaleurs de confiance) et le monde de la recherche. La société civile, de par sa proximité avec le terrain et sa compréhension fine de problématiques spécifiques (santé publique, lutte contre les discriminations…), est un maillon essentiel dans l'identification des sujets émergents qui nécessitent un éclairage, et facilitera le travail des chercheurs.

Il sera aussi opportun de réfléchir aux modalités du **partage de données** entre les signaleurs de confiance et le monde de la recherche.

[32] L'Arcom a déjà initié des échanges avec des chercheurs spécialisés dans l'étude des plateformes lors des deux campagnes électorales de 2022, qui ont nourri un rapport de l'institution : Rapport sur les campagnes électorales 2022 : élection à la présidence de la République et élections législatives | Arcom.

➢ *Développer des capacités d'analyse agiles*

Les fonctionnalités et les conditions d'utilisation des services déployés par les plateformes, ainsi que les risques qui peuvent émerger sur ces plateformes, évoluent de façon continue. Le régulateur doit donc pouvoir s'appuyer sur des **capacités d'analyse agiles**. Ces capacités, qui reposent sur l'identification préalable d'interlocuteurs dans l'écosystème de recherche, ont un intérêt particulier dans le cadre d'évènement majeurs (élections, Jeux Olympiques, …) ou d'annonces de la part des plateformes (changement de politique, mise en place de nouvelles fonctionnalités). Lors de telles situations, le régulateur doit pouvoir faciliter la circulation d'information permettant d'éclairer le débat et de guider l'action des pouvoirs publics.

➢ *Faciliter la hiérarchisation des besoins en nouvelles données et de mise à jour des API*

Il est important que les chercheurs identifient et communiquent leurs **besoins en nouvelles données et en matière d'intégration de nouvelles fonctionnalités par les API.** Cette consultation fait également ressortir la nécessité de **hiérarchiser** et **centraliser** ces besoins afin de présenter des demandes claires aux plateformes et de permettre un dialogue constructif avec ces dernières. Le développement de nouvelles fonctionnalités requiert des ressources humaines et financières, et peut poser des questions de protection des données ou de sécurité.

Ces réflexions, déjà amorcées par certains groupes de chercheurs[33], pourraient être structurées à l'échelle de chaque plateforme, dans la mesure où ces dernières disposent de fonctionnalités et de profils d'utilisateurs différents. L'Arcom travaille actuellement avec le PEReN pour mettre en place des formats d'échange avec les chercheurs sur ces sujets, et suit les initiatives en cours de déploiement par l'EDMO.

➢ *Protéger les accès existants et inciter la mise en place d'API*

Une grande majorité des chercheurs avec qui l'Arcom échange soulèvent l'actualité relative au changement de politique de Twitter concernant l'accès des chercheurs à son API, qui **met en péril de nombreux projets actuels et futurs**. De même, il est urgent qu'une plateforme comme TikTok, à la croissance rapide (notamment chez les jeunes) et au format novateur (basé largement sur la circulation virale de contenus vidéos de courte-durée), mette une API à disposition des chercheurs européens.

L'Arcom souhaite donc contribuer à la **protection des accès existants aux API** des plateformes pour les chercheurs et à inciter la mise en place de programmes dédiés aux chercheurs pour les plateformes ne l'ayant pas encore fait. Dans cette optique, l'Arcom souligne l'importance des conditions générales d'accès à ces API, qui peuvent parfois faire entrave à la faisabilité des recherches.

➢ *Défendre un accès aux données publiques pour les médias, vérificateurs d'informations et pour les associations de la société civile*

L'Arcom appelle à ce que les **médias, ONG et vérificateurs d'informations** puissent également disposer d'accès aux données publiquement accessibles des plateformes, selon des modalités permettant leurs travaux. En effet, ces acteurs pourraient tirer profit de ces accès pour générer de la connaissance dans une temporalité différente de la recherche académique, et donc de manière complémentaire à elle.

[33] Notamment le projet De Facto: DIGITAL PLATFORMS' GOVERNANCE.pdf (defacto-observatoire.fr).

➢ *Engager une réflexion autour du don de données et du scraping*

L'Arcom souhaite également engager un dialogue avec les chercheurs autour de sujets plus prospectifs, tels que le passage à l'échelle du **don de données** aux chercheurs (notamment par le biais de coopératives de données, qui sont mentionnées dans le *Data Governance Act*) et **la faisabilité des audits dits « adversariels »** (reposant par exemple sur l'automatisation de comptes).

b) Tirer pleinement parti des nouvelles possibilités d'accès aux données

b.1) Assurer la bonne circulation de l'information au niveau français

L'évolution du cadre règlementaire réorganise profondément les modalités d'accès aux données des plateformes. Dans ce contexte, le régulateur doit pouvoir exposer clairement les nouvelles possibilités d'accès à l'écosystème de recherche français, ainsi que les critères et procédures permettant d'en bénéficier.

Ce travail sera important pour fluidifier la mise en place des nouveaux mécanismes, et promouvoir leur utilisation par les chercheurs. Le régulateur devra notamment informer les chercheurs quant à la montée en puissance de l'implémentation du RSN, aussi bien en France que chez nos partenaires européens, et éviter que les chercheurs ne rencontrent des difficultés dans la formulation de leurs demandes.

➢ *Expliquer et promouvoir les nouveaux mécanismes d'accès aux données*

Le régulateur devra mettre des **ressources à la disposition des chercheurs**, afin de permettre à ces derniers de construire leurs projets en ayant une meilleure compréhension de la faisabilité des accès sur lesquels ils peuvent s'appuyer. Il parait notamment opportun d'énoncer clairement les catégories de données disponibles publiquement (notamment via les API des plateformes) et les données qui pourraient être obtenues uniquement sur la base d'un agrément.

Ce travail d'analyse et de clarification des textes a en partie déjà été effectué dans la partie 3)a) de cette synthèse, sur laquelle l'Arcom s'appuiera afin de publier des **fiches ressources** à disposition des chercheurs, qui seront publiées dans une section dédiée de son site internet. En complément, l'Arcom pourra répondre directement aux questions des chercheurs (via la mise en place d'un **point de contact dédié**) notamment afin de les accompagner dans la préparation de leurs demandes d'accès.

➢ *Fournir exceptionnellement un accompagnement à certains projets*

L'Arcom se propose également de fournir très exceptionnellement un accompagnement à la préparation de certaines demandes d'agrément présentant un fort intérêt pour la collectivité et étant soumis à des contraintes de temps importantes.

La CNIL a développé ces dernières années un **accompagnement juridique** et **technique**, mis en place pour les acteurs amenés à traiter des données personnelles, notamment via la publication de ressources en ligne. Le programme d'accompagnement renforcé d'entreprises[34], lancé par l'autorité en début d'année 2023, constitue un exemple prometteur afin de faciliter le respect du RPGD dans le cadre de projets impliquant le traitement de données sensibles. Ce type d'accompagnement personnalisé

[34] « Accompagnement renforcé » : la CNIL lance un nouveau dispositif innovant d'accompagnement | CNIL

nécessite un fort investissement de la part de l'accompagnant et de l'accompagné, et devrait donc rester exceptionnel.

➢ *Œuvrer à la transparence des processus d'accès*

La mise en place du mécanisme d'agrément défini dans l'article 40 du RSN nécessite une montée en puissance des régulateurs européens et notamment du CSN des pays d'établissement des VLOPSEs. De la même manière, la mise en place de l'agrément défini par le Code de bonnes pratiques contre la Désinformation sera liée à la création d'un tiers de confiance sur la base des travaux du groupe de travail de l'EDMO.

Il sera donc important pour l'Arcom de **suivre le déploiement gradué** de ces mécanismes, notamment en ayant une bonne connaissance des **délais** de délivrance de l'agrément, des **critères** concrets retenus pour évaluer les demandes[35] ainsi que les potentielles **objections** à l'agrément. L'Arcom partagera avec les chercheurs les informations recueillies. Ce travail de suivi et de partage sera important dans la mesure où les mécanismes d'application de l'article 40 du RSN devraient monter en puissance au cours du temps. En effet, les parties prenantes auront besoin d'un temps d'adaptation : définition et maitrise de processus techniques sécurisés, opérationnalisation du concept de risque systémique, élaboration de premiers jeux de données par les plateformes, etc.

b.2) S'inscrire dans une logique de « réseau de régulateurs »

La mise en œuvre du RSN requiert une intensification de la collaboration entre les régulateurs nationaux, et avec la Commission européenne. L'Arcom est déterminée à contribuer à la montée en puissance des processus d'accès aux données des plateformes, en capitalisant sur sa maitrise des enjeux et en s'appuyant sur ses échanges avec les différents acteurs français de la recherche et les autorités compétentes françaises, dont la CNIL. L'Arcom se positionnera comme un partenaire pour les autres CSN et pour la Commission européenne s'agissant de l'accès aux données à des fins de recherche.

L'Arcom s'appuie déjà sur ses nombreux échanges avec la communauté de recherche pour contribuer aux travaux de la Commission européenne. Ainsi, l'Arcom a contribué de façon publique[36] aux travaux de la Commission portant sur le futur acte délégué complétant l'article 40 du RSN.

➢ *Faciliter le travail d'agrément du CSN d'établissement des VLOPSEs*

Le RSN octroie au CSN français la possibilité de procéder à une **évaluation initiale** des demandes d'agrément venant de France, afin de faciliter le travail du CSN d'établissement. Le RSN précise néanmoins que les chercheurs français pourront également choisir de transmettre leur demande directement au CSN d'établissement des VLOPSEs. L'Arcom devra donc avoir une valeur ajoutée pour les chercheurs français désirant obtenir un agrément.

Le CSN français national doit pouvoir jouer pleinement ce rôle de facilitation du travail du CSN d'établissement de la VLOPSE, qui devra traiter un grand nombre de demandes d'agrément. Afin d'accélérer le traitement des demandes et d'éviter les aller-retours, l'Arcom pourra assister les chercheurs basés en France dans l'élaboration de leurs demandes, notamment en s'assurant que leurs dossiers sont dûment circonstanciés.

[35] Notamment l'appréciation qui sera faite de la notion de « risque systémique ».
[36] Réponse de l'Arcom à l'appel à contribution de la Commission européenne sur l'acte délégué complétant l'article 40 du RSN - Feedback from: Arcom (europa.eu).

Le CSN du pays dans lequel l'organisme de recherche des chercheurs est affilié disposera des expertises permettant d'évaluer efficacement le respect de certains des critères d'agrément définis dans l'article 40-8, notamment l'indépendance des chercheurs vis-à-vis d'intérêts commerciaux, leur affiliation à un organisme de recherche et la transparence de leurs financements.[37]

L'Arcom monte également en compétence en matière de sécurité et de confidentialité des données, de proportionnalité des accès, et de mise à disposition des résultats. Elle devra pour cela s'associer aux acteurs nationaux disposant d'une forte expertise en la matière, notamment la **CNIL**, qui accompagne déjà certains chercheurs[38], mais aussi les DPO des universités et les laboratoires de recherche eux-mêmes. Dans cette optique, l'Arcom a déjà engagé un projet d'agrément « à blanc » afin de saisir les implications pratiques concrètes d'une demande au sens de l'article 40 du DSA, et notamment sur les éléments permettant de satisfaire les critères stipulés par l'article 40-8.

A terme, il sera bon que les CSN travaillent de concert afin d'harmoniser leur évaluation des différents critères d'agrément. Il serait bon également que les CSN puissent collaborer à la mise en place d'un formulaire standardisé des critères à satisfaire, comportant des indications sur les documents permettant de justifier la conformité avec ces critères.[39]

➢ *Opérer la liaison entre le niveau national et européen*

L'Arcom devra **faire remonter au niveau européen les conclusions du dialogue** qui sera mis en place au niveau national avec les parties prenantes (chercheurs, DPO des universités, société civile, autorités compétentes). Ce dialogue portera entre autres sur la définition des problématiques de recherche (incluant la détection de sujets émergents), les besoins en nouvelles données et l'identification de mesures techniques de protection des données.

L'Arcom compte capitaliser sur ses relations avec les chercheurs afin de disposer d'informations à jour sur les projets de recherche en cours au niveau national ainsi que d'être rapidement informé **des nouvelles problématiques** identifiées par les chercheurs et la société civile (associations, vérificateurs d'information, journalistes). Cela permettra à l'Arcom de contribuer aux travaux de la Commission européenne et des autres régulateurs européens en matière de **détection** et de **réduction des risques systémiques**.

L'Arcom développe également une expertise permettant le suivi de **la mise à disposition et de l'évolution des API** de recherche mises à disposition par les opérateurs de plateformes. L'Arcom portera une attention particulière aux conditions d'utilisation de ces API, afin que celles-ci soient compatibles avec les impératifs des chercheurs. L'Arcom pourra aussi participer au débat au niveau européen sur la mise à disposition de certaines **nouvelles données**, notamment sur celles qui pourraient être considérées comme non-sensibles.

---

[37] L'ANR dispose également de connaissances pointues des modalités et des acteurs du financement.
[38] l'article 44-6 de la loi « Informatique et Libertés » prévoit que tout traitement nécessaire à la *recherche publique impliquant des données sensibles* au sens de l'article 9 du RGPD sous réserve que des motifs d'intérêt public important les rendent nécessaires *doit faire l'objet d'un avis préalable motivé et publié de la CNIL*. Cette formalité préalable, qui ne vaut pas « autorisation », a vocation à accompagner les chercheurs dans la mise en conformité à la réglementation à la protection des données de leurs travaux de recherche.
[39] En ce sens, les ressources mises à disposition par la CNIL afin de faciliter le travail des responsables de traitement des données semblent être une bonne inspiration : L'analyse d'impact relative à la protection des données (AIPD) | CNIL.

Plus généralement, l'Arcom devra mettre en avant les réussites identifiées au niveau national, notamment les outils mis à disposition des chercheurs, les **mesures techniques** identifiées pour **garantir la sécurité** des données, ou encore les innovations en matière de **réutilisation** et de **transparence** des données de recherche.

➢ *Suivre les travaux de constitution d'un tiers indépendant*

Le groupe de travail de l'EDMO sur l'accès des chercheurs aux données des plateformes s'attèle actuellement à la constitution d'un **tiers de confiance indépendant** en charge de superviser l'agrément des projets de recherche rentrant dans le cadre du Code de bonnes pratiques contre la Désinformation.
Ce tiers de confiance développe une expertise spécifique sur le respect du RGPD lors du partage de données à des fins de recherche, notamment en identifiant les solutions techniques permettant de sécuriser le partage de données. Le RSN, quant à lui, laisse la possibilité de faire appel à ce tiers indépendant dans l'examen des demandes d'agrément des chercheurs (Article 40-13).

L'Arcom suit donc attentivement la mise en place de ce tiers de confiance en liaison avec la CNIL, autorité compétente sur la question des données personnelles, et le projet **De Facto**, branche française de l'EDMO. L'Arcom prêtera également attention à la transparence et à l'indépendance de ce tiers, au respect de ses décisions par les plateformes, et aux discussions quant à sa potentielle implication dans l'agrément des chercheurs au sens du RSN.

Au-delà de son rôle dans l'agrément des chercheurs, le groupe de travail de l'EDMO propose également que le futur tiers supervise **les documents d'information** publiés par les plateformes, informe les chercheurs sur la **disponibilité des données** et répertorie les projets de recherche. L'EDMO travaille également à consulter les chercheurs sur leurs **besoins en nouvelles données** afin de présenter aux plateformes des listes claires et hiérarchisées. L'Arcom est pleinement disposée à apporter son expertise au tiers sur ces sujets.

Dans le cadre du Groupe des régulateurs européens des services de médias audiovisuels (ERGA), l'Arcom suit également l'évaluation du respect des engagements pris par les plateformes signataires du Code de bonnes pratiques contre la désinformation, dont un certain nombre portent sur le partage de nouvelles données et le soutien à la recherche. Par ailleurs, ce nouveau code poursuivant des objectifs et reposant sur des principes similaires à ceux du système français basé sur la loi du 22 décembre 2018 (engagements sur des moyens, mesure de l'effectivité par des indicateurs, axes similaires et association des différentes parties prenantes), l'Arcom est pleinement impliquée dans les groupes de travail de la *taskforce* de la Commission européenne qui a pour objectif de maintenir le code à jour et adapté à sa finalité, en lien avec le monde académique.

Références

Arcom, (2022). « Lutte contre la manipulation de l'information sur les plateformes en ligne. Bilan 2021 ».

Allcott, H., Braghieri, L., Eichmeyer, S., & Gentzkow, M. (2020). "The welfare effects of social media." *American Economic Review*, 110(3), 629-676.

Bursztyn, L., Egorov, G., Enikolopov, R., & Petrova, M. (2019). "Social media and xenophobia: evidence from Russia." *National Bureau of Economic Research Working Paper* No. w26567.

Edelson, Laura, Inge Graef, et Filippo Lancieri. (2023). « Access to data and algorithms: For an effective DMA and DSA implementation ».

EDMO's Working group on Platforms-to-researcher Data Access et George Washington University's Institute for Data, Democracy & Politics. (2022). « Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access ».

Fujiwara, T., Müller, K., & Schwarz, C. (2021). "The effect of social media on elections: Evidence from the United States." *National Bureau of Economic Research Working Paper* No. w28849.

Guhl, Jakob, Oliver Marsch, et Henry Tuck. (2022). *Researching the Evolving Online Ecosystem: Barriers, Methods and Future Challenges.* ISD.

Husovec, Martin. 2022. « Will the DSA work? » *Verfassungsblog: On Matters Constitutional*.

Levy, R. E. (2021). "Social media, news consumption, and polarization: Evidence from a field experiment." *American Economic Review*, 111(3), 831-870.

Lurie, Emma. (2023). « Comparing Platform Research API Requirements ». *Tech Policy Press*.

Vermeulen, Mathias. (2022). « Researcher Access to Platform Data: European Developments ». *Journal of Online Trust and Safety* 1(4).

Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). "Political effects of the internet and social media." *Annual Review of Economics*, 12, 415-438.