DMP du projet "FeedingBias: A multi-platform mixed-methods approach to news exposure on social media "

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "FeedingBias: A multi-platform mixed-methods approach to news exposure on social media "
Livrable	D03
Version	Version initiale
Objet/périmètre du plan	Ce DMP couvre toutes les données collectées dans le cadre du projet ANR "Feeding Bias". Ces données sont celles collectées dans le cadre des work packages WP2, WP3 et WP5. Elles seront exploitées dans le cadre des autres work packages WP2, WP4 et WP6.
Domaines de recherche (selon classification de l'OCDE)	Sociology, Political science, Computer and information sciences, Mathematics
Langue	fra
Date de création	2023-06-07
Date de dernière modification	2023-10-26

Renseignements sur le projet

Titre du projet	FeedingBias: A multi-platform mixed-methods approach to news exposure on social media
Acronyme	FeedingBias
Résumé	FeedingBias vise à fournir des données et des outils analytiques pour une meilleure compréhension de l'exposition à l'information sur les réseaux sociaux et des interactions des individus avec ces informations. Notre objectif est a) de comprendre l'influence des caractéristiques sociodémographiques des utilisateurs, de leur orientation politique et de leur niveau de conscience algorithmique sur leur exposition à l'information, et b) de mesurer la contribution de ces biais d'exposition et d'engagement sur la polarisation de la sphère publique moderne. FeedingBias sera basé sur le recrutement d'un large échantillon d'utilisateurs des médias sociaux répondant à une enquête par questionnaire ainsi que sur la collecte d'information sur l'usage de Twitter, Facebook et Youtube par ces utilisateurs. Toutes les données collectées dans le cadre de FeedingBias seront traitées selon les exigences du RGPD et le dispositif de recherche sera évalué par le Comité d'éthique de l'Université Grenoble Alpes.
Sources de financement	• Agence Nationale de la Recherche :
Date de début	2023-01
Date de fin	2027-01-31
Partenaires	• Pacte - Laboratoire de sciences sociales https://ror.org/026j45x50 • Grenoble Images Parole Signal Automatique https://ror.org/02wrme198 • Laboratoire d'Informatique de l'Ecole Polytechnique https://ror.org/04afed728 • GROUPE DE RECHERCHE SUR LES ENJEUX DE LA COMMUNICATION

Produits de recherche :

1. Default research output

Contributeurs

Nom	Affiliation	Rôles
Bastin Gilles	Pacte	• Coordinateur du projet • Responsable du plan de gestion de données
Pacouret Jérôme	Pacte	• Personne contact pour les données

Droits d'auteur :
Le(s) créateur(s) de ce plan accepte(nt) que tout ou partie de texte de ce plan soit

réutilisé et personnalisé si nécessaire pour un autre plan. Vous n’avez pas besoin de citer le(s) créateur(s) en tant que source. L’utilisation de toute partie de texte de ce plan n’implique pas que le(s) créateur(s) soutien(nen)t ou aient une quelconque relation avec votre projet ou votre soumission.

DMP du projet "FeedingBias: A multi-platform mixed-methods approach to news exposure on social media "

1. Description des données et collecte ou réutilisation de données existantes

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Jeu de données n°1 (JD1) : caractéristiques des médias français

Work package 2 : Mapping the sources of online information

Responsable : Oana Goga (LIX) & Emmanuel Marty (GRESEC)

Données produites : Il s'agit de collecter des informations sur les propriétés et publications en ligne d'environ 600 médias français (sites d'information associés ou non à des journaux imprimés et des chaines de radio et de télévision). Ces données caractérisent leurs comptes et stratégies de publication sur les réseaux sociaux numériques ; les informations qu'ils diffusent sur les réseaux sociaux ; leurs propriétés et stratégies économiques et éditoriales.

Sources, méthodes et logiciels de production de données nouvelles : Les membres du projet collectent des informations sur les médias français sur les sites internet de ces médias, sur quatre réseaux sociaux numériques (Facebook, Instagram, Twitter, YouTube), sur le site de l'Alliance pour les chiffres de la presse et des médias (ACPM), ainsi que sur le site data.culture.gouv.fr. Ces données sont récoltées par l'intermédiaire de navigateurs ordinaires (Firefox, etc.) et compilées à l'aide de tableurs courants (Open Office, Excel) et du site Framaforms.

Données existantes : Les noms de certains des médias étudiés ont été identifiés dans des publications existantes. La collecte de ces nouvelles données est toutefois nécessaire pour objectiver la nature et certains déterminants de l'offre d'information journalistique en France. Il s'agit d'être plus exhaustifs que d'autres recensions des médias français, tout en récoltant des informations précises sur leur caractéristiques économiques et professionnelles et leurs publications sur internet.

Documentation de la collecte et livrables : La provenance et le processus de collecte de ces données sont régulièrement documentés sous la forme de comptes rendus de réunions, d'un rapport sur la base de données constituée, d'un compte-rendu de workshop sur la collecte de données (livrable 11 ou D11), d'un guide de collecte des données (D13), ainsi que des rapports intermédiaire et final du projet (D06 et D07).

Jeu de données n°2 (JD2) : usages des réseaux sociaux de la population française

Work package 3 : Users questionnaires conception and deployment

Responsable : Gilles Bastin (Pacte) & Emmanuel Marty (GRESEC)

Données produites : Il s'agit de collecter, à l'échelle d'un échantillon de plusieurs milliers d'individus représentatifs de la population française, des informations précises sur leurs pratiques culturelles usages des réseaux sociaux ; points de vue sur l'information, la politique et les réseaux sociaux ; propriétés sociodémographiques. Il s'agit aussi de récolter les identifiants de volontaires sur quatre réseaux sociaux : Facebook, Instagram, Twitter, YouTube.

Sources, méthodes et logiciels de production de données nouvelles : Ces données sont collectées à l'aide de questionnaires administrés *via* internet par les membres du projet, ainsi qu'un institut de sondage. Dans la suite de ce document, on désigne comme "répondants volontaires" les individus ayant à la fois répondu au questionnaire et accepté de partager leurs identifiants sur au moins un des quatre réseaux étudiés (Facebook, Instagram, Twitter et YouTube). Les questionnaires sont administrés à l'aide du logiciel d'enquête en ligne LimeSurvey et de la plateforme Screen de la MSH Alpes.

Données existantes : Par comparaison avec les enquêtes existantes, les questionnaires administrés permettent de récolter des informations plus précises sur les usages des réseaux sociaux numériques. Les questionnaires permettent aussi de collecter des identifiants, dans la perspective de croiser de façon originale l'étude des pratiques et des caractéristiques sociodémographiques des répondants (telles que déclarées en réponse aux questionnaires) avec l'observation des informations auxquelles ils sont exposés sur les réseaux sociaux étudiés.

Documentation de la collecte et livrables : La provenance et le processus de collecte de ces données sont régulièrement documentés sous la forme de comptes rendus de réunions, de rapports sur l'administration de chaque questionnaire (D33 à 36), d'un compte-rendu de workshop sur la collecte de données (D11), d'un guide de collecte des données (D13), ainsi que des rapports intermédiaire et final du projet (D06 et D07).

Jeu de données n°3 (JD3) : exposition à l'information sur les réseaux sociaux numériques

Work package 5 : Data collection time-windows supervision

Responsable : Gilles Bastin (Pacte) & Oana Goga (LIX)

Données : les données renseignent l'exposition à l'information des répondants volontaire sur les réseaux sociaux, à travers des variables renseignant par exemple le nombre et les types de médias suivis, le nombre de publications potentiellement consommées *via* les réseaux sociaux, l'orientation politique et journalistique de ces médias et publications, etc.

Sources, méthodes et logiciels de production de données nouvelles : Sur la base de JD1 et JD2, les membres du projet collectent des données sur les informations publiées par les médias étudiés (JD1) et auxquelles sont exposés les répondants volontaires par l'intermédiaire de leurs comptes sur les réseaux sociaux. Ces données couvrent les publications de médias

pendant deux périodes de six semaines. Elles sont collectées par l'intermédiaire des API et de pages de Facebook, Instagram, Twitter, YouTube. Pour cela, l'équipe utilise les logiciels R et Python.

Données existantes ; Des enquêtes ont déjà observées la réception de l'information sur les réseaux sociaux, mais leurs données ne sont pas accessibles et le plus souvent limitées à un seul réseau. En outre, l'originalité de notre projet est de pouvoir croiser l'étude des informations circulant sur les réseaux sociaux avec des informations sur les pratiques et caractéristiques des utilisateurs de ces réseaux, récoltées par questionnaire. Cette méthodologie originale implique la production de nouvelles données.

Documentation de la collecte : La provenance et le processus de collecte de ces données sont régulièrement documentés sous la forme de comptes rendus de réunions, de rapports sur la collecte de données (D51 et D52), d'un compte-rendu de workshop sur la collecte de données (D11), d'un guide de collecte des données (D13), ainsi que des rapports intermédiaire et final du projet (D06 et D07).

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Nom du jeu de données	Typologie des données	Formats des données	Volume (ordre de grandeur)
Jeu de données n°1 (JD1) : caractéristiques des médias français	• liste des médias français étudiés (sources_list) • liste des variables utilisées pour caractériser les médias étudiés (sources_variables) • base de données des identifiants des médias étudiés sur les réseaux sociaux (sources_id) • base de données des caractéristiques économiques et professionnelles des médias étudiés (sources_df)	Les listes seront codées en format .txt, .pdf. Les base de données seront codées en format .csv Pour faciliter le partage et la réutilisation des données, deux formats utilisés (.txt et .csv) sont des formats standards et ouverts permettant l'interopérabilité. Pendant la collecte et la mise en forme des données, d'autres formats peuvent être utilisés (docx, .xlsx, .odt). Mais toutes les versions finales des données seront en format ouvert.	Les listes de médias et de variables n'excéderont pas quelques dizaines de Ko chacune. Les bases de données ne devraient pas excéder 2Mo (quelques centaines de lignes et quelques dizaines de colonnes)
Jeu de données n°2 (JD2) : questionnaires sur les usages des réseaux sociaux de la population française	• Un questionnaire sur des usages des réseaux sociaux, pratiques culturelles, caractéristiques sociodémographiques et identifiants sur Facebook, Instagram, Twitter et YouTube (q1) • Un questionnaire sur des perceptions de sujets d'actualité et participations à la médiation et au commentaire de ces actualités sur les réseaux sociaux (q2) • Deux bases de données compilant l'ensemble des données récoltées *via* l'administration de Q1 et Q2 (users_df1 et users_df2) (à l'exception des identifiants sur les réseaux sociaux) • Une base de données comprenant le pseudonyme ou identifiant de tous les répondants, ainsi que leurs identifiants sur les réseaux sociaux étudiés (users_id_df) • Des bases de données comprenant chacune une fraction des données de users_df1 et/ou users_df2. Ces bases serviront à faciliter l'exploitation des données (en rendant possible le travail sur de plus petites bases, des sous-	Q1 et Q2 seront codés et exportés du logiciel Limesurvey en format .pdf, .html, et .txt. users_df1, users_df2 et users_id_df seront codées en format .csv. Pour faciliter le partage et la réutilisation des données, les formats utilisés (.txt, .csv, .html et .txt) sont des formats standards et ouverts permettant l'interopérabilité. Lors de la préparation des questionnaires et de l'exploitation des données, les formats docx., doc., odt. et xlsx. pourront être provisoirement utilisés, en fonction des préférences de l'équipe de recherche pour tel ou tel logiciel. Mais toutes les versions finales des données seront en format ouvert.	Les questionnaires Q1 et Q2 ne devraient pas excéder quelques centaines de Ko chacun (10 à 20 pages) Les bases de données users_df1, users_df2 et users_id_df ne devraient pas excéder chacune plus de quelques dizaines de Mo (quelques milliers de lignes et quelques centaines de colonnes de texte)

	échantillons, des données renommées pour être plus lisibles, etc.)		
Jeu de données n°3 (JD3) : exposition à l'information sur les réseaux sociaux numériques	• des scripts de collectes de données *via* les sites et les API de Facebook, Instagram, Twitter et YouTube • des bases de données sur l'ensemble des publications des médias étudiés sur chacun de ces réseaux, pendant deux périodes de six semaines (P1 et P2) • des bases de données sur ces informations accessibles aux répondants volontaires *via* leurs comptes personnels sur Facebook, Instagram, Twitter et YouTube (abonnement aux médias étudiés, informations reçues en provenance des médias étudiés)	Les scripts seront codés en format .r, .txt, .py et .ipynb Les bases de données seront codées en format .csv Pour faciliter le partage et la réutilisation des données, les formats utilisés pour les versions finales des documents sont des formats standards et ouverts permettant l'interopérabilité.	Les données collectées pourront représenter jusqu'à quelques centaines de Go

---

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

DOCUMENTATION ET METADONNEES

Jeux de données

Un document de type "lisez-moi" présentera les fichiers composant chaque jeu de données (JD1, JD2 et JD3) en indiquant :

- le nom du fichier
- le format du fichier
- le responsable de la production de ces données
- une description du type de données
- les conditions d'accès aux données

A l'issue du projet, une synthèse méthodologique réunira le compte-rendu du workshop consacré à la méthodologie, ainsi que les autres présentations et livrables consacrés à la méthodologie globale de l'enquête.

Fichiers de données

JD1 : données sur les médias français

- liste des médias étudiés (sources_list)
  - les principes de sélection de ces médias sont explicités dans un document à part, en format txt, et seront également publiés dans un data paper (avec DOI)
- liste des variables (sources_variables)
  - un livre de code précisant la signification de chaque variable et de chaque modalité
- base de données des identifiants (sources_id)
  - un document de type "lisez-moi" sur la collecte de ces identifiants (date, sources, problèmes rencontrés)
- base de données sur les caractéristiques des médias (sources_df)
  - un document de type lisez-moi sur la collecte des données (date, sources, problèmes rencontrés)
  - un data paper sur le choix et les caractéristiques des médias étudiés
  - une liste des publications présentant ces données et basées sur celles-ci (avec leurs DOI)

JD2 : données sur les usages des réseaux sociaux

- questionnaire n°1 (q1)
  - un fichier "lisez-moi" précisant le responsable et les auteurs du questionnaire, les dates de préparation et

d'administration, ainsi que les publications présentant le questionnaire et s'appuyant sur celui-ci
- questionnaire n°2 (q1)
    - un fichier "lisez-moi" précisant le responsable et les auteurs du questionnaire, les dates de préparation et d'administration, ainsi que les publications présentant le questionnaire et s'appuyant sur celui-ci (avec leurs DOI)
- base de données n°1 (users_df1)
    - un fichier lisez-moi précisant le responsable et les participants à l'administration du questionnaire, la date de passation, les autres métadonnées disponibles
    - un dictionnaire de code précisant la signification de chaque variable et de chaque modalité, ainsi que les questions du questionnaire ayant permis la récolte des données
    - le rapport de passation du questionnaire
    - Liste des publications tirées de la base (avec leurs DOI)
- base de données n°2 (users_df2)
    - un fichier "lisez-moi" précisant le responsable et les participants à l'administration du questionnaire, la date de passation, les autres métadonnées disponibles
    - un livre de code précisant la signification de chaque variable et de chaque modalité, ainsi que les questions du questionnaire ayant permis la récolte des données
    - le rapport de passation du questionnaire
    - liste des publications tirées de la base et précisant les méthodes et le processus de collecte des données (avec leurs DOI)

JD3 : informations circulant sur les réseaux sociaux

- scripts de collecte de données
    - un fichier de type lisez-moi pour chaque script, précisant l'auteur du script, la date d'écriture, la source des données qu'il permet de collecter et les données collectées
- bases de données sur les informations circulant sur chaque plateforme
    - pour chaque base, un fichier de type "lisez-moi" précisant le responsable de la collecte, le type de données collectées, la période de collecte, les autres métadonnées disponibles, ainsi que les autres données et métadonnées utilisées pour la construction de cette base (scripts et données de JD1 et JD2)
    - un livre de code précisant la signification de chaque variable et de chaque modalité, ainsi que les questions du questionnaire ayant permis la récolte des données
    - Liste des publications tirées de la base et précisant les méthodes et le processus de collecte des données (avec leurs DOI)

QUALITE DES DONNEES

Organisation des données

Pendant l'enquête, sur l'intranet du projet, les données et documents de travail sont organisées par work package. Dans les dossiers consacrés à chaque workpackage, les données et métadonnées à jour sont stockées dans un dossier à part. Les autres dossiers réuniront respectivement les documents utilisés pour le travail en cours (bases de données provisoires, textes en cours de rédaction) et les archives.

A l'issue de l'enquête, les données et métadonnées finales seront réorganisées dans trois dossiers correspondant chacun à un jeu de données.

Règles de nommage des fichiers

Les noms des versions à jour des fichiers de données sont définis dès le début de l'enquête (sources_list, users_df1, etc.)

Les fichiers sont nommés selon les recommandations de DoRANum, visant à intégrer les principes FAIR :

- pas d'espace, de caractère spécial ou de majuscule
- date au format AAAAMMJJ
- format type : type_version_date_dernierscontributeurs
- ex : sources_id_V1-20230303_EM

FORMATION, PRODUCTION ET OUTIL DES METADONNEES

Une session du workshop méthodologique sera consacrée à la définition et la production des métadonnées, ainsi qu'à la mutualisation des savoir-faire et bonnes pratiques à ce sujet.

Au fil de l'enquête, les participants rapprocheront leurs pratiques des standards DDI, qui est le standard universel le plus commun pour les sciences sociales, et du Dublin Core, notamment via l'usage de l'interface "Dublin Core Generator", qui est d'accès relativement aisé pour les profanes. (https://nsteffel.github.io/dublin_core_generator/generator_nq.html). Sur la compatibilité des standards DDI et du Dublin Core, voir https://ddialliance.org/resources/ddi-profiles/dc.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Mesures de contrôle transversales

Des rendez-vous réguliers entre les responsables de chaque work package et l'ensemble de l'équipe permettent de coordonner la construction des trois jeux de données, et ainsi de s'assurer de la bonne articulation des méthodes et

disciplines mobilisées par l'enquête. Ces rendez-vous sont des occasions de présentation des méthodes et bonnes pratiques propre à chaque discipline et équipe associée au projet.

Un workshop méthodologique est organisée pendant la première année du projet de façon à rendre compte et discuter des méthodes et outils utilisées, en présence de chercheurs et ingénieurs compétents extérieurs à l'équipe du projet ANR.

Des publications méthodologiques sont préparées par toute l'équipe, dont au moins un data paper sur la recension des médias français diffusant leurs informations sur les réseaux sociaux.

Mesures de contrôle propres à chaque jeu de données

JD1 : données sur les médias français

Les médias sont identifiés en croisant des sources officielles (Commission paritaire des publications et agences de presse) et les listes de médias constituées par des équipes d'experts reconnus (Medialab, projet ANR PIL, etc.). Les médias étudiés sont retenus en fonction de critères de sélection explicites et définis et justifiés en référence aux travaux théoriques et méthodologiques existants. Les choix opérés et les résultats seront présentés dans un article méthodologique soumis à une revue à comité de lecture spécialisée dans la méthodologie.

JD2 : données sur les usages des réseaux sociaux

Préparation des questionnaires : Les questionnaires sont préparés en référence aux questionnaires d'enquêtes nationales sur les pratiques culturelles et numériques des français, ainsi qu'en référence à la littérature multidisciplinaire sur les pratiques culturelles et les usages des réseaux sociaux. Avant administration, les questionnaires sont soumis à l'examen de professionnels spécialisés d'un ou de plusieurs instituts de sondage couramment sollicités par des organismes de recherche publique, ainsi qu'au délégué à la protection des données de l'UGA.

Administration : Le questionnaire n°1 (q1) est administré auprès d'un échantillon représentatif de la population adulte française établi à travers la méthode des quotas et constitué par un institut de sondage compétent en matière de sous-traitance d'enquêtes sociologiques par questionnaire. Après contrôle de la qualité des données par l'institut de sondage, l'équipe du projet ANR en charge de ce jeu de données se chargent d'un nouveau contrôle consistant à éliminer les questionnaires mal ou peu remplis.

Traitement : Toutes les bases de données constituées précisent les sous-échantillons qu'elles agrègent, de façon à pouvoir comparer les conditions de constitution de ces sous-échantillons. En préalable de l'analyse poussée des résultats, les données sont comparées à des données de même type constituées par d'autres organisations ou équipes de recherche.

Cette méthodologie est évaluée via des communications dans des événements scientifiques et la publication d'articles dans des revues à comité de lecture. Ces publications seront rendues accessibles *via* les métadonnées et la documentation (voir 2.a).

JD3 : informations circulant sur les réseaux sociaux

Les scripts de collecte sont constitués conformément aux normes scientifiques et aux conditions d'utilisation des réseaux sociaux étudiés.

Les bases de données sont constituées et évaluées en comparaison avec d'autres recherches scientifiques de même type, sur la base des publications tirées de ces dernières.

Cette méthodologie est évaluée via des communications dans des événements scientifiques et la publication d'articles dans des revues à comité de lecture. Ces publications seront rendues accessibles via les métadonnées et la documentation (voir 2.a).

---

3. Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Les données sont stockées en respectant les recommandations de DoRANum et tout particulièrement la règle 3-2-1 :

- à distance sur la plateforme Cloud UGA administrée par l'UGA. Ce service dispose d'une fonction "versions précédentes" limitant les risques de suppression ou de dégradation involontaire des fichiers.
- les bases de données volumineuses seront également stockées à distance sur les serveurs de l'Unité d'appui à la recherche nommée Grenoble Alpes Recherche Infrastructure de Calcul Intensif et de Données, (GRICAD, ou UAR 3759, sous la tutelle du CNRS et de l'UGA).
- Si le volume des données excèdent la capacité de stockage offerte par Cloud UGA, et/ou pour accroitre encore la sécurité des données, nous aurons également recours à une autre plateforme administrée par l'UGA ou aux solutions offertes par la Plateforme universitaire de données Grenoble Alpes (PUD-GA). Ces plateformes alternatives sont présentées sur cette page : https://scienceouverte.univ-grenoble-alpes.fr/donnees/stocker/plateformes-locales/ et https://www.msh-alpes.fr/plateformes/pud-ga
- sur l'ordinateur professionnel du responsable du projet ANR. Les fichiers seront synchronisés en continu depuis et vers Cloud UGA, grâce à l'application NextCloud. Cet ordinateur sera chiffré à l'aide d'un logiciel comme BitLocker ou VeraCrypt.
- sur un disque dur externe conservé dans un autre bâtiment que celui où est localisé l'ordinateur professionnel du responsable du projet (sauvegarde tous les trois mois minimum). Ce disque dur ne sera pas connecté au réseau internet (les sauvegardes régulières devront donc être faites hors ligne). Ce disque dur sera chiffré à l'aide d'un

logiciel comme BitLocker ou VeraCrypt.

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Protection des données sensibles

Les données sensibles collectées sont des données personnelles concernant l'état civil, la vie personnelle, la vie professionnelle, les revenus, les données de connexion, les opinions politiques, la participation politique récente.

Les membres du projet n'auront accès qu'à des données pseudonymisées. Les données divulguées hors de l'équipe seront pseudonymisées.

Le responsable du projet sera le seul à avoir accès aux données de connexion des répondants au questionnaire, qui peuvent permettre leur identification. A cette fin, ces données seront stockées dans un fichier chiffré et protégé par un mot de passe dont il sera le seul détenteur - sur un ordinateur et un disque dur chiffrés.

Le ou les sous-traitants chargés de la passation du questionnaire n'auront pas accès aux données collectées lors de l'enquête, qui seront stockées directement sur les serveurs de la MSH-Alpes avant d'être transférées sur les espaces de stockage du projet (voir 3a).

Les principes de la procédure de pseudonymisation :

- non-accès de l'équipe aux noms et autres données directement identifiantes des enquêtés (prénom, nom, date de naissance, adresse IP, etc.), qui ne sont accessibles qu'à l'institut de sondage fournissant des données pseudonymisées (c'est à dire un jeu de données où chaque individu est désigné par un chiffre)
- seul le responsable du projet aura accès aux données dont les recoupements présentent un risque modéré ou élevé d'identification des personnes : identifiants sur les réseaux sociaux et code postal
- les données diffusées hors de l'équipe permettront des analyses au niveau individuel, mais ne comprendront pas les variables présentant un risque élevé, moyen ou modeste d'identification des personnes : identifiants sur les réseaux sociaux, code postal, comptes suivis sur les réseaux sociaux

Niveaux d'accès aux données des membres du projet :

De façon à protéger les données sensibles et à contrôler la bonne organisation et documentation des données produites, quatre niveaux distincts d'accès aux données (ou profils d'habilitation) sont attribués aux responsables et membres du projet :

- niveau 1 : le responsable du projet ANR a accès à toutes les données produites et stockées dans le cadre du projet. Il est le seul à avoir accès aux données pouvant permettre d'identifier relativement aisément les répondants aux questionnaires (c'est à dire leurs identifiants sur les réseaux sociaux étudiés, récoltés via q1, ainsi que leur code postal). Il est également le seul à pouvoir accéder au stockage des données sur les serveurs du GRICAD, sur son ordinateur personnel et sur le disque dur externe.
- niveau 2 : Ce niveau désigne les droits d'accès des membres de l'équipe listés dans le projet ANR et des postdoctorants. Ces derniers ont accès à toutes les données stockées sur Cloud UGA et ont le droit d'éditer ces données et d'en déposer de nouvelles sur cette même plateforme. Contrairement au responsable du projet, ils n'ont accès qu'à des données pseudonymisées.
- niveau 3 : Ce niveau est celui des stagiaires et d'autres contributeurs ponctuels. Ils ont accès à Cloud UGA en lecture seule. Les données qu'ils produiront seront déposées sur Cloud UGA par l'intermédiaires de participants de niveau 1 ou 2.
- niveau 4 : Pendant toute la durée du projet, les personnes extérieures consultées pour leur expertise scientifique ou pour faciliter la conduite de l'enquête n'ont pas accès à Cloud UGA et aux données brutes produites sans être accompagnées par des membres du projet de niveau 1, 2 ou 3.

Les permissions d'accès obsolètes seront supprimées tous les six mois.

Sécurisation des postes de travail

Les postes de travail de tous les membres du projet sont fournis par l'UGA, protégés par un mot de passe et un verrouillage automatique de session, et équipés d'un antivirus à jour et d'un pare feu logiciel.

En cas d'incident

Les données pourront être récupérées grâce à l'existence de plusieurs sauvegardes à distance et de plusieurs sauvegardes sur les disques durs des membres du projet (voir 3a.).

En cas de violation des données personnelles, le responsable de l'équipe se tournera immédiatement vers les services de l'UGA et du CNRS compétents pour être orienté sur la procédure à suivre. L'incident sera déclaré à la CNIL dans les 72 heures.

---

4. Exigences légales et éthiques, codes de conduite

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Nature des données personnelles collectées :

Les données à caractère personnel collectées concernent l'état civil, la vie personnelle, la vie professionnelle, les revenus, les données de connexion, les opinions politiques, la participation politique récente.

Identification et choix des données personnelles collectées :

Dès la préparation des questionnaires q1 et q2, les membres du projet identifient, parmi les informations ils prévoient de collecter, celles qui constituent des données à caractère personnel. Selon le premier "grand principe du RGPD" ("ne collectez que les données vraiment nécessaires"), seules seront collectées des données personnelles permettant d'analyser scientifiquement l'exposition à l'information sur internet , selon les critères de scientificité propres aux communautés disciplinaires des membres du projet.

Consentement et collecte des données :

Les questionnaires permettront d'obtenir le consentement éclairé des répondants pour la récolte de données à caractère personnel *via* le questionnaire et *via* l'observation de leurs comptes sur les réseaux sociaux étudiés (Facebook, Instagram, Twitter et YouTube). Ce consentement sera obtenu conformément aux deuxième et troisièmes "grand principes du RGPD" : la transparence et la facilitation de l'exercice des droits des personnes sur leurs données.

- le questionnaire q1 propose plusieurs consentements :
  - à la conservation et à l'utilisation des données pseudonymisées constituées par les réponses au questionnaire
  - à l'observation des données publiques sur les comptes d'utilisateurs de chaque réseau. En d'autres termes, les personnes ayant consenti à remplir le questionnaire auront ensuite la possibilité de consentir ou non à l'observation de chacune des plateformes qu'ils utilisent.
- le caractère éclairé de ce consentement et l'exercice des droits des répondants sont garantis par :
  - en préambule du questionnaire, un texte résumant les finalités de collecte des données à caractère personnel, la nature de ces données, les organismes de recherche utilisateurs de ces données, leur durée de conservation, les possibilité de réutilisation des données anonymisées par d'autres recherches, la non-utilisation à des fins commerciales, et l'anonymisation des données.
  - un lien vers le site internet du projet présentant plus en détail les finalités du projet, les données collectées, leur utilisation et l'équipe du projet.
  - une adresse email permettant de contacter le responsable du traitement des données
  - une adresse email permettant de contacter le délégué à la protection des données du CNRS en cas de problème concernant cette enquête

Pseudonymisation des données

- non-accès de l'équipe aux noms et autres données directement identifiantes des enquêtés (prénom, nom, date de naissance, adresse IP, etc.), qui ne sont accessibles qu'à l'institut de sondage fournissant des données pseudonymisées (c'est à dire un jeu de données où chaque individu est désigné par un chiffre)
- seul le responsable du projet aura accès aux données dont les recoupements présentent un risque modéré ou élevé d'identification des personnes : identifiants sur les réseaux sociaux et code postal
- les données diffusées hors de l'équipe permettront des analyses au niveau individuel, mais ne comprendront pas les variables présentant un risque élevé, moyen ou modeste d'identification des personnes : identifiants sur les réseaux sociaux, code postal, comptes suivis sur les réseaux sociaux

Durée limitée de conservation des données ("grand principe du RGPD")

- 6 mois après récolte des données sur les réseaux sociaux dans le cas des données de connexion (identifiants sur les réseaux sociaux), dans la limite de trois ans après recueil du consentement
- 5 ans après le recueil des données et du consentement dans le cas des autres données à caractère personnel

Sécurisation des données ("grand principe du RGPD")

voir 3b.

Mise en conformité

Les processus de collecte et de traitement de données personnelles seront déclarés auprès du relais DPO de l'UGA et de la déléguée à la protection des données du CNRS.

La conformité au RGPD sera également vérifiée à l'occasion des révisions de ce PGD.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Les membres de l'équipe s'engagent à respecter la législation en vigueur tout au long du projet, en matière de RGPD et de protection des droits de propriété intellectuelle notamment.

Les bases de données partagées au delà de l'équipe seront l'objet d'une licence de type Creative Commons choisie de façon à protéger le droit de paternité des créateurs et créatrices des bases de données tout en interdisant toute utilisation commerciale de ces bases et des oeuvres dérivées de ces bases.

Ces données seront ouvertes au plus tard cinq ans après collecte ou deux ans après la fin de financement du projet par l'ANR. Voir 5a.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Tout au long du projet, ses membres doivent s'interroger sur les questions éthiques liées à la collecte et au traitement de données personnelles et de données sur l'utilisation des réseaux sociaux. Ils doivent tout particulièrement réfléchir collectivement aux conditions d'un consentement éclairé pour des données récoltées en ligne, auprès de plateformes aux conditions d'utilisation obscures pour la grande majorité de leurs utilisateurs.

A cette fin, les membres du projet prévoient, dans le cadre de leurs rendez-vous réguliers et du workshop méthodologique, des temps de discussion des questions éthiques. Leurs réflexions s'appuient sur la littérature académique consacrée au relations entre les plateformes, la production et l'exploitation de données et la dissémination de la connaissance et de la méconnaissance de ces processus parmi les enquêtés potentiels.

---

5. Partage des données et conservation à long terme

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Jeu de données	Données	Services de diffusion	Moment de partage et embargo éventuel	Justification de l'ouverture et de la fermeture selon le principe "aussi ouvert que possible, aussi fermé que nécessaire"
Jeu de données n°1 (JD1) : caractéristiques des médias français	Liste et identifiants des médias étudiés sur les réseaux sociaux (sources_list et sources_ID)	• Site internet du projet • Nakala, un entrepôt de données de recherche pour les sciences humaines et sociales (Infrastructure de recherche "étoile" Human-Num IR*, mise en oeuvre par le CNRS avec le Campus Condorcet et Aix-Marseille université) • et/ou Progedo, un autre entrepôt et service de diffusion de données en sciences humaines et sociales, qui permet leur inscription au catalogue européen du CESSDA (Infrastructure de recherche "étoile", mis en oeuvre par le CNRS)	Ouverture dès publication d'un *data paper* sur les médias étudiés	Une recension étendue et scientifiquement contrôlée des médias français en activité sur les réseaux sociaux, tout comme leurs identifiants sur ces réseaux, constituent des données utiles à des recherches très variées sur les médias et les réseaux sociaux. La communauté scientifique bénéficierait d'un accès rapide à ces données, qui perdent en qualité relativement vite (en raison de la disparition et de la création fréquentes de médias).

Jeu de données n°1 (JD1) : caractéristiques des médias français	Base de données des caractéristiques économiques et professionnelles des médias étudiés (sources_df)	• Site internet du projet • Nakala, un entrepôt de données de recherche pour les sciences humaines et sociales (Infrastructure de recherche "étoile" Human-Num IR*, mise en oeuvre par le CNRS avec le Campus Condorcet et Aix-Marseille université) • et/ou Progedo, un autre entrepôt et service de diffusion de données en sciences humaines et sociales, qui permet leur inscription au catalogue européen du CESSDA (Infrastructure de recherche "étoile", mis en oeuvre par le CNRS)	Ouverture dès publication d'un article de synthèse sur l'offre d'information accessible *via* les réseaux sociaux numériques	Plusieurs communautés disciplinaires peuvent tirer profit d'une base de données riche sur les caractéristiques des médias français, et cela au service de questions de recherche variées. En partageant ces données après publication d'un article de synthèse, l'équipe se réserve la possibilité de rentabiliser l'investissement professionnel que constitue la production de ces données.
Jeu de données n°2 (JD2) : questionnaires sur les usages des réseaux sociaux de la population française	Questionnaires (q1 et q2)	Site internet du projet Archive ouverte HAL et/ou Nakala ou PROGEDO, en accompagnement des bases de données sur les usages des réseaux sociaux (voir ci-dessous)	Dès après la collecte des données	Le partage rapide des questionnaires est utile à l'intéressement de la communauté scientifique à notre enquête. Il peut être utile en vue d'enquêtes par questionnaires menées par d'autres chercheurs, chercheuses et équipes.
Jeu de données n°2 (JD2) : questionnaires sur les usages des réseaux sociaux de la population française	Bases de données sur les usages des réseaux sociaux de la population française et des répondants volontaires (users_df1 et users_df2)	• Site internet du projet • Nakala, un entrepôt de données de recherche pour les sciences humaines et sociales (Infrastructure de recherche "étoile" Human-Num IR*, mise en oeuvre par le CNRS avec le Campus Condorcet et Aix-Marseille université) • et/ou Progedo, un autre entrepôt et service de diffusion de données en sciences humaines et sociales, qui permet leur inscription au catalogue européen du CESSDA (Infrastructure de recherche "étoile",	Cinq ans après la collecte des données ou deux ans après la fin de financement du projet ANR	Ces données peuvent être utiles à la communauté scientifique pour mener des recherches ou enseigner sur les pratiques numériques de la population. L'embargo permet de rentabiliser l'investissement professionnel que constitue la production de ces données.

		mis en oeuvre par le CNRS)		
Jeu de données n°2 (JD2) : questionnaires sur les usages des réseaux sociaux de la population française	Identifiants des répondants sur les réseaux sociaux	Non partagé	Non partagé	La fermeture est nécessaire car ces données rendraient possible l'identification des répondants.
Jeu de données n°3 (JD3) : exposition à l'information sur les réseaux sociaux numériques	Scripts	• Archive ouverte HAL en cas de data paper sur la collecte de données sur les réseaux sociaux • Software Heritage si les scripts sont suffisamment originaux pour intéresser la communauté informatique	Dès publication éventuelle d'un data paper sur la collecte de données sur les réseaux sociaux	Données utiles à la conduite d'enquête nouvelles sur les réseaux sociaux
Jeu de données n°3 (JD3) : exposition à l'information sur les réseaux sociaux numériques	Bases de données sur les informations diffusées par les médias français sur les réseaux sociaux	• Nakala, un entrepôt de données de recherche pour les sciences humaines et sociales (Infrastructure de recherche "étoile" Human-Num IR*, mise en oeuvre par le CNRS avec le Campus Condorcet et Aix-Marseille université) • et/ou Progedo, un autre entrepôt et service de diffusion de données en sciences humaines et sociales, qui permet leur inscription au catalogue européen du CESSDA (Infrastructure de recherche "étoile", mis en oeuvre par le CNRS)	Cinq ans après la collecte des données ou deux ans après la fin de financement du projet ANR	Ces données peuvent être utiles à la communauté scientifique pour l'étude de l'offre et de la médiation des informations journalistiques en France. L'embargo permet de rentabiliser l'investissement professionnel que constitue la production de ces données.
Jeu de données n°3 (JD3) : exposition à l'information sur les réseaux sociaux numériques	Bases de données sur l'exposition des individus volontaires aux informations sur les réseaux sociaux	Non partagé	Non partagé	La fermeture est nécessaire car ces données pourraient permettre l'identification des participants, tout en renseignant sur leurs opinions et pratiques politiques

---

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Voir 5a pour le détail des données préservées pendant une durée indéterminée *via* leur dépôt sur l'archive ouverte HAL et les entrepôts de données en sciences humaines et sociales (Nakala et Progedo).

Utilisations possibles des données partagées : recherches et enseignements de sciences sociales et d'informatique sur les médias, les réseaux sociaux, la médiation et la réception d'information journalistique, les pratiques numériques et culturelles de la population française, les rapports de classe, de genre et de génération, les idées politiques des années 2020, etc.

La préservation des données plus de dix ans après leur partage n'est pas nécessaire en raison du présentisme de la grande majorité des recherches sur les pratiques numériques et la circulation d'information sur internet.

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Les données sont gérées selon les principes FAIR et seront donc faciles à trouver, accessibles, interopérables et réutilisables. En particulier, les formats utilisés sont des formats ouverts favorisant l'interopérabilité (voir 1b). Les données seront donc aisément accessibles à l'aide d'un ordinateur équipé d'un système d'exploitation, d'un navigateur et de logiciels open source de traitement de données.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Les plateformes et entrepôts de données utilisés pour conserver et partager les données, c'est à dire HAL, Nakala, Progedo et Software Heritage, proposent toutes des identifiants uniques et pérennes reconnus dans leurs domaines (DOI).

---

6. Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Le responsable de la gestion des données de l'ensemble du projet est le coordinateur du projet ANR, Gilles Bastin (Pacte). Il est responsable de la mise en oeuvre du PGD, de sa mise à jour et de sa révision.

Les co-responsables de la production, du stockage et du partage des jeux de données sont :

- Jeu de données 1 : Oana Goga (LIX) & Emmanuel Marty (GRESEC)
- Jeu de donnés 2 : Gilles Bastin (Pacte) & Emmanuel Marty (GRESEC)
- Jeu de données 3 : Gilles Bastin (Pacte) & Oana Goga (LIX)

Le draft du premier PGD a été préparé par Jérôme Pacouret (Pacte).

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Une session du workshop méthodologique sera consacrée à la bonne application de ce plan de gestion des données.

Ce plan sera révisé en concertation avec les membres du projet à l'occasion de la préparation du rapport intermédiaire du projet.

La révision et l'application du PGD, et tout particulièrement le stockage et le partage des données, pourront constituer des tâches constitutives du poste d'un ou de plusieurs des postdocs membres du projet.