FeedingBias_media : une base de données en accès libre de 894 médias français d'information générale actifs sur les réseaux socio-numériques

Jérôme Pacouret, Emmanuel Marty, Emma Orsolini, Gilles Bastin

FeedingBias_media : une base de données en accès libre de 894 médias français d'information générale actifs sur les réseaux socio-numériques

Auteurs et contacts
Jérôme Pacouret est docteur en sociologie et postdoctorant à Pacte (Sciences Po Grenoble)
pacouret.jerome@gmail.com

Emmanuel Marty est maitre de conférences en sciences de l'information et de la communication, membre du GRESEC (Université Grenoble Alpes)
emmanuel.marty@univ-grenoble-alpes.fr

Emma Orsolini est doctorante en sciences de l'information et de la communication, membre du GRESEC (Université Grenoble Alpes)
emmaorsolini@orange.fr

Gilles Bastin est professeur de sociologie à Sciences Po Grenoble (Pacte). Il dirige le projet ANR Feeding Bias.
gilles.bastin@sciencespo-grenoble.fr

Résumé
Ce data paper présente une base de données en accès libre permettant d'identifier et de comparer 894 médias français d'information générale qui diffusent leurs informations sur Facebook, Instagram, X (ex-Twitter) et/ou YouTube. Cette base de données offre un recensement actualisé et étendu des médias actifs sur quatre réseaux socio-numériques (RSN) parmi les plus utilisés pour s'informer en France. Le data paper explicite et détaille les principes et opérations de recensement des médias : en sélectionnant des médias reconnus comme tels par diverses institutions centrales du champ journalistique, aux visions contrastées de ce qu'est un média, notre approche dépasse les limites de l'autodéfinition des médias et de définitions institutionnelles trop restrictives. La base de données comprend les adresses et identifiants permettant d'accéder aux sites et comptes des médias étudiés, mais aussi des données sur leur ancienneté, leur volume de production et leur visibilité sur Facebook, Instagram, et YouTube. Ces données peuvent servir des recherches très variées, qu'elles portent sur un ou plusieurs RSN, et sur la production, la circulation ou la réception des informations.

Mots clés
médias, information générale, politique, Facebook, YouTube, Instagram, Twitter

Lien vers la base de données
https://nakala.fr/10.34847/nkl.cbfa30ni

Ce data paper présente une base de données en accès libre permettant d'identifier et de comparer 894 médias français d'information générale qui diffusent leurs informations sur Facebook, Instagram, X (ex-Twitter) et/ou YouTube[1]. La centralité de ces réseaux socio-numériques (RSN) dans la médiation et l'économie des informations, les nombreux problèmes publics qui leur sont associés et les données massives et extrêmement diversifiées qu'elles offrent aux chercheur·euses se sont traduits par une multiplication des recherches sur l'information en ligne. Parmi ces travaux très variés, plusieurs projets collectifs se sont déjà donnés pour but d'éclairer l'offre d'information journalistique accessible en ligne et sur les RSN en prenant pour objet plusieurs dizaines ou centaines de médias (Cagé *et al.*, 2017 ; Cointet *et al.* 2021 ; Lyubareva et Marty 2022 ; Marty *et al.*, 2012). Mais en raison des problématiques et approches spécifiques à chacun de ces projets et des transformations rapides de l'offre d'information en ligne, le long et complexe travail de recensement des médias à étudier est difficilement cumulatif. D'autant plus que les formats courants de publication scientifique laissent généralement peu de place à la description des critères de sélection et des méthodes d'identification des médias pris pour objet. Or, faut-il le rappeler, ce travail de recensement, dont la difficulté tient notamment à la coexistence de nombreuses définitions concurrentes des médias et du journalisme, conditionne les questionnements et les résultats des recherches sur l'information en ligne.

Ce data paper apporte trois contributions principales au travail d'identification et d'analyse des médias. La première est un recensement actualisé et étendu des médias d'information générale, qui couvrent l'actualité politique, et qui rendent accessibles leurs informations sur quatre RSN parmi les plus utilisés pour s'informer en France. Les données partagées ici permettent notamment des analyses s'étendant à plusieurs RSN, qui sont encore très rares alors qu'une large fraction de la population accède désormais à l'actualité en utilisant plusieurs de ces réseaux (Pacouret *et al.* 2024). Deuxièmement, le data paper explicite et détaille les principes et opérations de sélection des médias étudiés. Ces principes de sélection peuvent être appropriés et complétés par des recherches aux objets et approches théoriques variés : en sélectionnant des médias reconnus comme tels par diverses institutions centrales du champ journalistique, aux visions contrastées de ce qu'est un média, notre approche dépasse les limites de l'autodéfinition des médias et de définitions institutionnelles trop restrictives. Troisièmement, la base de données comprend les adresses et identifiants permettant d'accéder aux sites et comptes des médias étudiés, mais aussi des données sur leur ancienneté, leur volume de production et leur visibilité sur Facebook, Instagram, X (ex-Twitter) et YouTube. Décrites ici, ces données renseignent les inégalités de productivité et de visibilité entre les médias identifiés, tout en facilitant la constitution de sous-échantillons de médias à étudier.

Ces données et le data paper qui les accompagne peuvent donc être utiles à des recherches très variées, qu'elles soient qualitatives ou quantitatives, qu'elles construisent de grands ou de petits échantillons (en filtrant les médias les plus productifs ou visibles par exemple), et qu'elles portent sur la production, la circulation ou la réception des informations. La première section de cet article présente le processus de définition des médias étudiés et de collecte des données. La deuxième section présente les données partagées, en listant les variables de la base, puis en décrivant l'ancienneté, le multipositionnement sur les RSN, le volume et le rythme de publication, ainsi que le nombre d'abonnés et de vues des médias étudiés.

[1] Cette base de données a été construite dans le cadre du projet Feeding Bias, financé par l'ANR (ANR-22-CE38-0017-01). Elle est accessible à l'adresse https://nakala.fr/10.34847/nkl.cbfa30ni

1. Le processus de recensement des médias

1.1. Le cadre scientifique et institutionnel de création de la base de données : le projet Feeding Bias

La création de cette base de données s'inscrit dans le projet multidisciplinaire *Feeding Bias*, soutenu par l'Agence nationale de la recherche (ANR) après avoir été porté par la chaire Société Algorithmique du Multidisciplinary Institute for Artificial Intelligence (MIAI, Université Grenoble Alpes). Ce projet mobilise des chercheur·euses en sociologie, sciences de l'information et de la communication, informatique et mathématiques appliquées. Il analyse les déterminants sociaux de l'exposition à l'information et de l'engagement sur les RSN, notamment pour mieux objectiver des logiques d'exposition sélective, incidente et algorithmique aux informations, ainsi que des inégalités sociales d'accès à l'information et des inégalités de visibilité entre les médias. L'originalité de ce projet repose sur une approche mixte en termes de méthodes : des données sur les pratiques numériques d'utilisateurs des RSN, récoltées en ligne auprès d'un échantillon de volontaires, sont croisées avec un questionnaire renseignant leurs propriétés sociales et leurs pratiques numériques et culturelles.

L'un des volets du projet, dont est tiré ce data paper, est plus spécifiquement tourné vers l'analyse du paysage contemporain de l'offre d'information en ligne – cette analyse étant nécessaire à celle de l'exposition de la population à des médias variés. Douze ans après une précédente catégorisation complète des sources d'information en ligne (Marty et al., 2012), le paysage médiatique en ligne a considérablement évolué. Certains médias, comme les blogs, ont progressivement perdu une partie de leur influence, tandis que des *pure players* sont devenus centraux à la faveur du développement des réseaux socio-numériques tels que Facebook, YouTube, Instagram et Twitter, qui occupent aujourd'hui une position dominante dans l'économie politique de l'information en ligne (Smyrnaios et Rebillard, 2019 ; Sebbah et al., 2020 ; Mattelart, 2020). En outre, les anciennes catégorisations, qui distinguaient les médias traditionnels et les *pure players*, la presse écrite, l'audiovisuel et le web, tendent à perdre de leur pertinence à mesure que les frontières s'estompent entre les médias en raison de leur hybridation (Cagé et al., 2017). Les frontières et équilibres politiques du champ journalistique ont aussi évolué avec l'apparition de médias d'extrême droite de forte ou moyenne audience, comme CNews ou TVLibertés, mais aussi de nouveaux médias faiblement dotés en ressources et/ou à la lisière de la légitimité journalistique, qui se disputent souvent le label de l'indépendance (Sedel 2021). Enfin, la création de nouveaux médias est désormais une finalité et un instrument de la politique du ministère de la Culture, qui subventionne plus d'une dizaine d'incubateurs et d'autres dispositifs de soutien à l'entrepreneuriat journalistique. S'il faut se garder de croire à un bouleversement du champ journalistique – qui reste largement dominé par des médias dits traditionnels – les quelques transformations évoquées suffisent à rappeler que le recensement des médias est complexe et doit être fréquemment actualisé.

1.2.La définition des médias étudiés

Les médias inclus dans la base de données se définissent comme des sources d'information générale, qui couvrent l'actualité politique française, qui sont reconnues comme journalistiques par différentes entités (voir *infra*), et qui diffusent une fraction au moins de leurs productions sur des réseaux sociaux très utilisés en France pour diffuser et consommer des informations (Facebook, Instagram, YouTube et X ex-Twitter).

Les médias ont été définis et identifiés à l'aide de quatre critères. Le premier est la p*résence sur les RSN* : chaque média possède au moins un compte sur Facebook, Instagram, X ex-Twitter et/ou YouTube, où sont diffusées des informations. Le deuxième est la *reconnaissance par les institutions journalistiques*. Les médias sont reconnus comme des

médias ou des sources d'information légitimes par des institutions centrales du champ journalistique : médias, syndicats patronaux, administrations publiques et centres de recherche (voir 1.3). Le troisième critère est *la couverture de l'actualité politique*, au sens large. Les médias étudiés consacrent ainsi tout ou partie de leur activité à couvrir les institutions politiques, les professionnels de la politique, les mouvements sociaux, les idées politiques, et les enjeux de luttes sociales politisées. Enfin, le dernier critère est *territorial* : les médias étudiés sont établis en France et consacrent une part significative de leurs informations à l'actualité politique française (y compris à un échelon infranational).

La base de données a été constituée de façon à inclure le plus grand nombre possible de médias correspondant à ces critères de sélection, c'est-à-dire selon une logique d'exhaustivité. Cette base de données ne comprend pas pour autant toutes les sources d'information sur l'actualité française, loin de là. Les critères de sélection impliquent en effet d'exclure les médias inactifs sur les réseaux considérés ou ne traitant pas l'actualité politique (du fait d'une très forte spécialisation thématique, par exemple), ainsi que les sources d'informations non reconnues comme journalistiques par les institutions centrales de ce champ.

1.3.Le recensement des médias de la base de données

La base de données a été constituée entre octobre 2021 et juillet 2024. Le recensement a été opéré de façon à maximiser le nombre et la diversité des médias répondant aux critères de sélection présentés précédemment. À cette fin, nous nous sommes appuyés sur diverses sources, qui se distinguent par leurs critères de définition des médias d'information, mais qui ont toutes en commun d'être produites par des institutions du champ journalistique dotées d'un pouvoir de prescription de ce qu'est un média. Ainsi, notre recensement, tout en écartant les organisations se définissant comme des médias sans être reconnues comme telles, tient néanmoins compte du fait que le champ journalistique se caractérise par la coexistence d'institutions en concurrence pour la définition de ce qu'est un média, c'est-à-dire pour la définition des frontières du champ.

Un premier ensemble de sources comprend les *sources administratives*. Il s'agit d'une part de la liste de 207 médias dont les sites sont les plus fréquentés en France selon l'association Alliance pour les chiffres de la presse et des médias (ACPM, classement unifié d'octobre 2021). Et d'autre part de la liste des services de presse reconnus par la Commission paritaire des publications et agences de presse (CPPAP) et plus spécifiquement, la sous-liste de 363 médias officiellement reconnus comme des services de presse en ligne d'information politique et générale au sens de l'article 2 du décret du 29 octobre 2009[2], ainsi que la sous-liste de 199 services de presse en ligne consacrés pour une large part à l'information politique et générale au sens de l'article 39bis A du code général des impôts[3]. Le classement de l'ACPM regroupe et hiérarchise, selon des critères d'audience, des médias reconnus comme tels par des dirigeants d'entreprises de presse, puisque le conseil d'administration de cette alliance est composé de « 26 éditeurs représentant les grandes organisations professionnelles de la Presse, 4 annonceurs et agences et 2 administrateurs désignés ». Quant à la CPPAP, elle réunit à égalité des représentants syndicaux des entreprises de presse et des représentants des ministères de la

[2] Cet article dispose que « sont considérés comme d'information politique et générale les services de presse en ligne dont l'objet principal est d'apporter, de façon permanente et continue, des informations, des analyses et des commentaires sur l'actualité politique et générale locale, nationale ou internationale susceptibles d'éclairer le jugement des citoyens. Ces informations doivent présenter un intérêt dépassant significativement les préoccupations d'une catégorie de lecteurs. L'équipe rédactionnelle doit comporter au moins un journaliste professionnel, au sens de l'article L. 7111-3 du code du travail. »

[3] Les critères de reconnaissance de cet article sont : éditer une publication de périodicité au maximum mensuelle ou un service de presse en ligne ; apporter de façon permanente sur l'actualité politique et générale, locale, nationale ou internationale des informations et commentaires tendant à éclairer le jugement des citoyens ; consacrer au moins le tiers de leur surface rédactionnelle à cet objet.

Communication, de la Culture, du Budget et de l'Economie. Les médias qu'elle reconnait comme consacrés en totalité ou pour une large part à l'informations politique et générale ont en commun de consacrer au moins un tiers de leurs productions à ce type de sujets et de satisfaire plusieurs critères de professionnalisation journalistique : comprendre au moins un journaliste professionnel, adopter une périodicité régulière, respecter les obligations de la loi de 1881 sur la liberté de la presse, ne pas consacrer plus des deux tiers de leur surface à la publicité et aux annonces et servir « l'intérêt général »[4].

Le deuxième ensemble de sources est constitué de *sources académiques* (qui loin de seulement décrire le champ journalistique de l'extérieur, contribuent à la définition de ses frontières à travers leur travail de définition et d'analyse des médias et non-médias). Ces sources sont les listes de médias étudiés par quatre projets de recherche sur l'offre d'information en ligne : la liste de 199 médias du projet ANR « Internet, pluralisme et redondance de l'information – IPRI » ; la liste de 87 médias du projet OTMédia+ (INA) ; la liste de 420 médias du projet « Media Polarization à la française » (Institut Montaigne, Medialab et école de journalisme de Sciences Po, MIT Center for Civic Media) ; la liste de 84 médias du projet ANR « Pluralisme de l'Information en ligne ». Ces quatre listes et projets avaient en commun de proposer une vue d'ensemble de l'offre d'information en ligne en France, mais au service de problématiques variées, comme l'évolution des modèles d'affaires des médias, le pluralisme de l'agenda médiatique ou la polarisation politique de l'information. Ce faisant, ces projets ont cherché à diversifier les médias inclus dans leurs échantillons au regard des supports de diffusion (*pure players*, sites associés à des journaux et chaînes de télévision et de radio), des modèles d'affaires (publicité, abonnement, don, etc.), des orientations politiques (du centre à des médias d'extrême gauche et d'extrême droite), de l'échelle territoriale (médias locaux, régionaux et nationaux) et de la légitimité auprès des institutions centrales du champ journalistique (notamment car la liste du Medialab est basée sur une liste établie par le service de fact-checking du *Monde*, qui comprend aussi bien des médias dominants que des sites labellisés comme de « réinformation » et disqualifiés comme producteurs de désinformation).

Un troisième ensemble de sources comprend *les sources professionnelles sur les médias situés aux périphéries du champ journalistique*. Ces sources ont été mobilisées de façon à mieux représenter dans la base de données les médias se définissant comme « indépendants » ou « militants », de gauche radicale, d'extrême droite, de création récente, et/ou excessivement accessibles *via* les RSN et d'autres « plateformes » numériques. Pour intégrer des médias se définissant comme « indépendants » et « libres », et pour la plupart situés à la gauche du champ journalistique, les sources employées sont une liste des 138 sites internet accessibles via le « portail des médias libres »[5] et la liste des 100 médias et collectifs à l'initiative des Etats généraux de la presse indépendante du 30 novembre 2023[6]. Pour mieux couvrir l'extrême droite, aux médias déjà recensés par le Medialab ont été ajoutés quelques médias identifiés dans le cadre d'une recherche de notre équipe sur *Valeurs actuelles* (Marty et al. 2022). Par ailleurs, des médias de création récente et valorisés comme innovants ont été identifiés parmi 403 entreprises soutenues par quatorze incubateurs de nouveaux médias et entreprises culturelles[7], ainsi que parmi quelques dizaines de médias valorisés dans une newsletter sur l'innovation des

[4] Voir le site de la CPPAP, http://www.cppap.fr/criteres-dadmission/

[5] Ce portail est administré par Bastamag et propose officiellement « un accès direct à l'ensemble de la presse indépendante et aux articles publiés par les médias de « transformation » sociale, écologique et démocratique ». Ces médias ont été recensés en 2022. Voir https://portail.basta.media/spip.php?page=sources

[6] Voir le site du Fonds pour une presse libre, https://fondspresselibre.org/liberons-linfo-30-nov

[7] Les incubateurs consultés ont pour la plupart été identifiés en tant que bénéficiaires d'une subvention du ministère de la Culture. Une minorité seulement des entreprises qu'ils soutiennent sont des médias couvrant l'actualité politique et générale. Ces incubateurs sont YouM3edia, The Media House, le Medialab de TF1, Nm Cube (Ouest Media Lab), Mediastart, Media Lab 93, Le Tank Media, La Rotative, Creatis, Belle de Mai, Hôtel 71, Le Mas, Off7, Les Echos/Le Parisien.

médias (Médianes). Enfin, 26 médias aux informations essentiellement diffusées *via* les RSN et d'autres plateformes numériques ont été identifiés parmi les participants au Paris Podcast Festival (compétition catégorie documentaire, entre 2018 et 2023) et dans des articles de presse consacrés à des podcasts et à des vidéastes couvrant l'actualité *via* YouTube, TikTok et Twitch[8].

L'exclusion des médias ne couvrant pas l'actualité politique a été opérée au fil du recensement des médias, sur la base des pages d'accueil de ces sites et de leurs comptes sur les réseaux sociaux.

1.4. La collecte des données sur les comptes Facebook, Instagram, YouTube et X ex-Twitter

La collecte des identifiants des médias sur les réseaux sociaux (URL et/ou nom de compte ou de chaîne) a été opérée parallèlement au recensement des médias étudiés, jusqu'en février 2024. Ces identifiants ont été recherchés d'abord sur les sites internet des médias, qui comprennent généralement des liens vers leurs comptes sur les réseaux sociaux, puis par l'intermédiaire du moteur de recherche Google si le site du média étudié ne renvoyait pas vers l'un ou l'autre des RSN étudiés. Ces recherches ont permis d'exclure de la base de données des médias inactifs depuis trois ans au moins. En revanche, nous avons inclus à la base les identifiants de comptes inactifs de médias qui poursuivent leur activité en employant d'autres moyens de diffusion. Ces comptes pourraient de nouveau être utilisés à l'avenir et leurs identifiants permettent d'enquêter sur des périodes plus anciennes.

La collecte des identifiants, et plus généralement l'analyse de l'activité des médias sur les RSN, sont complexifiées par le fait que nombre de médias étudiés possèdent plusieurs comptes tel ou tel réseau. Des comptes spécialisés peuvent par exemple être dédiés à une rubrique particulière, ou à un sous-lectorat, comme dans le cas du *Monde* qui possède des comptes Twitter spécialisés dans l'actualité politique ou sur l'Afrique, ainsi qu'un compte à destination du public anglophone. Expérimentée dans le cas de 50 médias à forte audience, l'identification de ces comptes secondaires montre que leur nombre est très variable selon les médias et peut s'élever à plus de dix par RSN, mais aussi que beaucoup de ces comptes secondaires sont désormais inactifs et/ou republiaient des informations déjà publiées par les comptes principaux des médias. Pour limiter la redondance des données collectées et faciliter la comparaison des médias étudiés, la collecte a finalement été restreinte aux seuls comptes dits principaux des médias sur chaque RSN, que nous définissons comme ceux qui possèdent le plus grand nombre d'abonnés (sauf s'il s'agit du compte personnel d'un journaliste)[9].

Une autre difficulté tient au fait que les médias d'un même groupe, tout en possédant des sites internet distincts, partagent parfois un même compte sur un RSN. C'est le cas de médias du groupe Presse et Médias du Sud-Ouest, ainsi que d'autres médias locaux. Pour pouvoir étudier le plus grand nombre de médias possible, nous avons fait le choix de caractériser tous les médias recensés par leurs comptes, quitte à ce que certains de ces comptes figurent en double (ou triple). Ces doublons peuvent être aisément identifiés et exclus de l'analyse si nécessaire.

Enfin, les données sur l'activité des comptes ont été récoltées en juin 2024 *via* le service Apify, qui permet de collecter des données *via* les API des RSN étudiés. Ce choix a été fait en raison de la facilité d'utilisation et du coût réduit de ce service, dans un contexte de fermeture

---

[8] Bodgan Bodnar, « Marre des chaînes d'infos ? Voici 10 chaînes YouTube, Twitch, TikTok pour suivre la présidentielle jusqu'au second tour », *Numerama,* 22 avril 2022 ; Nino Barbey, « Vulgarisation politique, impression 3D, théâtre : 4 comptes TikTok à suivre en juillet 2023 », *Numerama*, 1er juillet 2023 ; « Ces youtubeurs qui font dans la politique », *L'ADN*, 25 avril 2017 ; Marin Tézenas du Montcel, « de gauche à droite, les chaînes politiques prolifèrent sur les réseaux » ; 30 septembre 2021 ; Pauline Demange-Dilasser, « Cinq chaînes YouTube pour s'informer différemment », *Télérama*, 30 avril 2023.

[9] Dans la très grande majorité des cas, il s'agit du compte généraliste du média, qui porte son nom et a vocation à présenter l'ensemble de ses informations. Mais dans le cas de C8, il s'agit du compte de l'émission la plus populaire de la chaîne, TPMP.

relative de l'accès des chercheur·euses à des API comme ceux de X (ex-Twitter)[10]. En raison de différences entre les RSN et leurs API, les données ne sont pas toujours équivalentes pour chaque réseau. La date de création des comptes n'a pas pu être récoltée dans le cas d'Instagram, et le volume de publications dans le cas de Facebook. Et YouTube est le seul réseau pour laquelle nous disposons du nombre de vues des publications.

2. Description de la base de données et de l'activité des médias sur les RSN

Cette section présente le contenu de la base de données, puis de premiers traitements statistiques de celle-ci qui rendent compte de la diversité des médias inclus dans la base en termes de positionnement, d'ancienneté, de volume de publication et de visibilité sur Facebook, Instagram, YouTube et X.

2.1. Métadonnées de la base feedingbias_media

La base de données, intitulée feedingbias_media, est accessible au format CSV, choisi car ouvert et d'utilisation très courante. Ces données sont conservées et partagées sur Nakala, qui est un entrepôt national de référence pour les données des sciences humaines et sociales. La première colonne de cette base comprend l'identifiant unique du média qui lui a été attribué lors de la récolte des données. Les 28 autres colonnes comprennent les données détaillées dans le tableau 1.

Tableau 1. Contenu de la base de données

Groupes de données	Noms des variables	Contenu des variables
Identifiant dans la base	media_id	Numéro identifiant le média dans la base de données
Adresses et identifiants sur les RSN	url	URL de la page d'accueil du site internet du média
	youtube_id	Identifiant de la chaîne YouTube
	youtube_url	URL de la chaîne YouTube
	x_id	Identifiant du compte X
	x_url	URL du compte X
	facebook_url	URL de la page Facebook
	instagram_url	URL du compte Instagram
Publications et visibilité sur Instagram	instagram_use	Variable dichotomique indiquant la possession d'un compte Instagram. Elle a été codée à l'aide des autres données récoltées.
	instagram_name	Nom du compte Instagram
	instagram_posts	Nombre de posts du compte Instagram
	instagram_followers	Nombre d'abonnés du compte Instagram
Publications et visibilité sur Facebook	facebook_use	Variable dichotomique indiquant la possession d'un compte Facebook. Elle a été codée à l'aide des autres données récoltées.

[10] Le 9 février 2023, X ex-Twitter a mis fin à la gratuité de son API au profit d'un abonnement mensuel allant de 100$ à 5000$ selon les formules, supprimant dans le même mouvement l'API « Twitter Academics ». La recherche universitaire s'en est trouvée largement empêchée, passant de la possibilité de collecter gratuitement des millions de tweets par jour à la nécessité de payer 100$ mensuels pour 10 000 tweets sur la période. Voir https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

	facebook_name	Nom du compte Facebook
	facebook_creation	Date de création du compte Facebook
	facebook_followers	Nombre d'abonnés du compte Facebook
	facebook_likes	Nombre de likes du compte Facebook
Publications et visibilité sur YouTube	youtube_use	Variable dichotomique indiquant la possession d'une chaîne YouTube. Elle a été codée à l'aide des autres données récoltées.
	youtube_name	Nom de la chaîne YouTube
	youtube_creation	Date de création de la chaîne YouTube
	youtube_followers	Nombre d'abonnés de la chaîne YouTube
	youtube_videos	Nombre de vidéos de la chaîne YouTube
	youtube_views	Nombre total de vues de la chaîne YouTube
Publications et visibilité sur X (ex-Twitter)	x_use	Utilisation ou non-utilisation de X
	x_name	Variable dichotomique indiquant la possession d'un compte X. Elle a été codée à l'aide des autres données récoltées.
	x_creation	Date de création du compte X
	x_posts_classic	Nombre de posts classiques du compte X
	x_posts_media	Nombre de posts de type "médias" du compte X
	x_followers	Nombre d'abonnés du compte X

2.2.Positionnement et ancienneté des médias sur les quatre RSN étudiés

La base comprend 894 médias présents sur au moins l'un des quatre réseaux étudiés (voir le tableau 2). Leur présence est maximale sur X ex-Twitter (91,8% des médias ont un compte) et minimale sur YouTube (67,1% des médias ont une chaîne). Chaque RSN compte au moins 600 des médias étudiés.

Tableau 2. Présence des médias sur les RSN

RSN utilisés	Nombre de médias	%
Médias possédant un compte sur Facebook	796	89,0%
Médias possédant un compte sur Instagram	636	71,1%
Médias possédant une chaîne sur YouTube	600	67,1%
Médias possédant un compte sur X (ex-Twitter)	821	91,8%
Nombre total de médias de la base	**894**	**100,0%**

La très grande majorité des médias étudiés sont présents sur plusieurs RSN (voir le tableau 3). Seulement 65 (7,3%) ne sont présents que sur l'un d'eux. Près de la moitié des médias sont présent sur Facebook, X, YouTube et Instagram (431 médias, soit 48,2% de l'échantillon).

Tableau 3. Multipositionnalité des médias sur les RSN

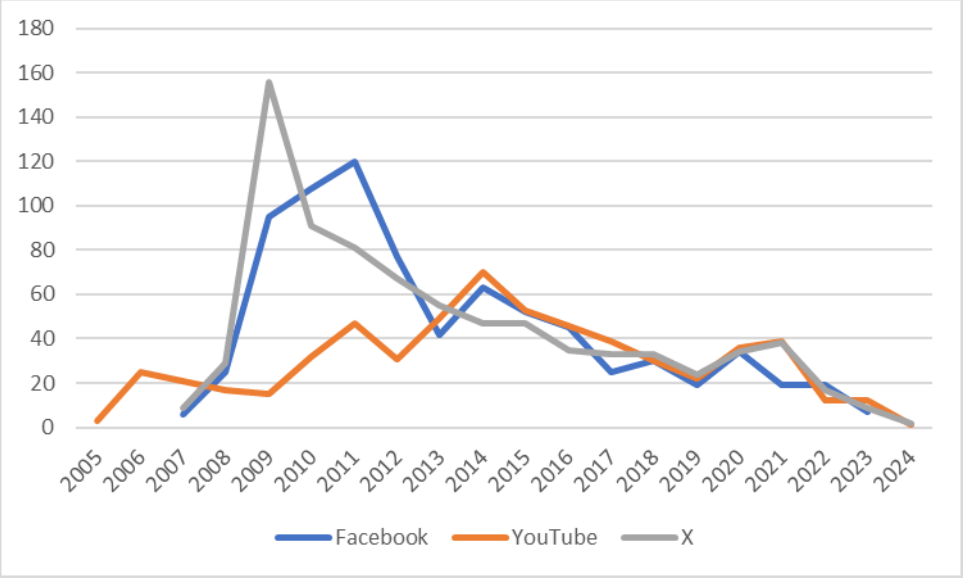
Nombre de RSN utilisés par les médias (à l'exclusion de tout autre réseau que Facebook, Instagram, X et YouTube seulement)	Nombre de médias	%
4 réseaux	431	48,2%

3 réseaux	268	30,0%
2 réseaux	130	14,5%
1 réseau	65	7,3%

La base comprend des médias d'ancienneté très variable. Bon nombre d'entre eux sont des médias dits traditionnels, dont l'existence précède généralement l'émergence des RSN et même d'internet. Il s'agit de journaux, de chaînes de radio et de chaînes de télévision qui ont étendu leur diffusion à internet et aux RSN. La base comprend aussi des médias présents exclusivement en ligne, dits *pure players*, créés avant ou parallèlement au développement des RSN. Enfin, on y trouve des médias qui diffusent principalement ou exclusivement leurs productions *via* les RSN et d'autres plateformes, c'est-à-dire des médias dits « natifs des réseaux sociaux ». Hugo Décrypte est l'un des plus connus des médias de ce type, qui comptent aussi des chaînes YouTube d'extrême droite par exemple.

L'ancienneté des médias sur chaque RSN est variable (voir le graphique 1). Les créations de comptes Twitter et Facebook par des médias ont connu un pic au tournant des années 2010, qui traduit l'investissement par les médias traditionnels de ces services alors en cours de massification. Dans le cas de YouTube, cette croissance fut plus progressive et connut un pic au milieu des années 2010, au moment où se généralisait le format vidéo sur différents réseaux (notamment Facebook) et dans des médias jusque-là spécialisés dans l'information écrite. Passées ces périodes de forte croissance, des comptes ont continué à être créés, à un rythme plus lent, notamment par de nouveaux médias.

Graphique 1. Création de comptes Facebook, YouTube et X par année



2.3.Volumes et rythmes de publication : grands et petits producteurs d'information

La base comprend aussi bien les plus grands producteurs d'information sur l'actualité française que de très petits médias. Les médias de l'échantillon se distinguent par le volume et la fréquence de leurs publications sur chaque RSN étudié, ce qu'on ne peut objectiver que dans le cas d'Instagram, YouTube et X. Depuis la création de leurs comptes, les médias étudiés ont publié entre 0 message ou vidéo et plusieurs dizaines ou centaines de milliers (voir le tableau 4). Une fraction importante des médias utilisant chaque réseau y ont très peu publié : 15,3% ont

publié moins de 1 000 tweets, 21% moins de 100 messages sur Instagram, et 43% moins de 100 vidéos sur YouTube (voir les tableaux 5, 6, 7). La base comprend aussi des médias qui publient une grande quantité de messages et de vidéos, et qui à eux seuls représentent une grande proportion des publications de l'ensemble des médias de la base. Ainsi, le décile des médias qui publie le plus sur chaque réseau totalise au moins la moitié des publications de l'ensemble des médias de la base sur le même réseau (voir le tableau 4). En d'autres termes, la production d'information sur les différents RSN étudiés est très concentrée autour de quelques médias qui produisent beaucoup plus que les autres.

Si Instagram, YouTube et X voient coexister de petits et très petits producteurs d'informations et de grands et très grands producteurs d'information, ces trois réseaux se distinguent par le volume et le rythme des publications des comptes (voir le tableau 4). Le volume de production est minimal dans le cas d'Instagram et YouTube, et maximal et très supérieur dans le cas de X. Ces variations tiennent en partie au coût variable des formats de publication : les photos diffusées sur Instagram et les vidéos diffusées sur YouTube sont bien plus coûteuses à produire que les tweets (au moins pour les médias qui produisent leurs propres images). Mais ces variations tiennent aussi aux formats et aux usages différenciés des RSN : Twitter sert beaucoup plus à la diffusion et au commentaire de l'actualité chaude, dans une logique de flux. Enfin, le volume relativement faible de messages diffusés sur Instagram s'explique en partie par la création et la massification plus récentes de ce réseau.

Outre le volume des publications, c'est leur concentration entre les médias qui varie selon les réseaux. Cette concentration est maximale dans le cas de YouTube, ou le décile des chaînes les plus productrices publient 84,8% des vidéos publiées par l'ensemble des chaînes de l'échantillon (contre 59,2% dans le cas de X et 52,5% dans le cas d'Instagram).

Tableau 4. Volume de production sur Instagram, YouTube et X

Nombre de publications	Instagram (posts)	YouTube (vidéos)	X (tweets)
Minimal	0	1	0
Maximal	18 696	138 453	1 158 263
Moyen	1 544	2 220	46 315
Médian	557	157	10 315
Proportion des publications publiées par le décile de compte publiant le plus sur le réseau considéré	52,5%	84,8%	59,2%

Tableau 5. Nombre de publications sur Instagram

Nombre de post	Médias	Proportion des comptes Instagram de la base
Moins de 100	130	21,2%
100 à 1 000	261	42,5%
1 000 à 5 000	167	27,2%
Plus de 5 000	56	9,1%

Tableau 6. Nombre de vidéos sur YouTube

Nombre de vidéo	Médias	Proportion des comptes de la base
Moins de 100	261	43,5%
100 à 1 000	209	34,8%
Plus de 1 000	130	21,7%

Tableau 7. Nombre de tweets

Nombre de tweets	Médias	Proportion des comptes de la base
Moins de 1 000	124	15,3%
1 000 à 10 000	277	34,3%
10 000 à 100 000	299	37,0%
Plus de 100 000	108	13,4%

La comparaison des volumes totaux de production de chaque compte est un indicateur imparfait de la fréquence de publication, puisque ces comptes sont plus ou moins anciens. Dans le cas de X ex-Twitter au moins, on peut vérifier que les médias se distinguent fortement par la fréquence de leurs publications. Près d'un tiers des comptes sur ce réseau y ont publié moins de 400 messages par an (soit à peine plus d'un par jour). À l'inverse, près d'un dixième des comptes ont publié plus de 10 000 tweets par an (soit plusieurs dizaines de tweets par jour) (voir le tableau 8).

Tableau 8. Nombre de tweets par an

Nombre de tweets	Médias	Proportion des comptes de la base
Moins de 400	242	30,1%
De 400 à 1000	149	18,5%
de 1000 à 10 000	341	42,4%
Plus de 10 000	73	9,1%

2.4.L'inégale visibilité des médias : nombre de followers et nombre de vues

Les médias de l'échantillon sont très inégalement visibles sur les réseaux sociaux. On peut le mesurer d'abord au nombre d'abonnés à leurs comptes. Pour chaque réseau, le nombre minimal est nul et le nombre maximal est compris entre 10 et 30 millions (voir le tableau 9). Une grande proportion des médias de la base sont très peu visibles. Un quart des comptes sur chaque réseau ne comptent que quelques centaines ou milliers d'abonnés (voir les tableaux 10, 11, 12, 13). Et la majorité des médias de la base n'y sont visibles que par une toute petite fraction de la population : la médiane du nombre d'abonnés varie de 5 111 à 23 305 selon les RSN. En revanche, entre 14,6% et 27,2% des médias présents sur chaque réseau y comptent au moins 100 000 abonnés. Plus encore que le volume de production, la visibilité des informations est donc très concentrée au profit d'une petite fraction des médias.

La visibilité des médias varie selon les RSN. Elle est maximale dans le cas de Facebook et minimale dans le cas d'Instagram (voir le tableau 9). Le nombre moyen et le nombre maximal de followers sont ainsi trois fois supérieur dans le cas de Facebook que dans celui d'Instagram. Cela s'explique en partie par le nombre d'utilisateurs de chaque réseau, qui est beaucoup plus élevé dans le cas de Facebook que dans celui d'Instagram.

Tableau 9. Nombre d'abonnés sur les quatre plateformes

Nombre d'abonnés	Instagram	Facebook	YouTube	X(ex-Twitter)
Minimum	0	0	1	0
Maximum	10 119 758	30 006 764	18 100 000	10 872 206
Moyenne	107 306	337 406	152 966	159 609
Médiane	5 111	23 305	1 400	7 984

Tableau 10. Nombre d'abonnés des médias sur Instagram

Nombre d'abonnés	Nombre de médias	Proportion des médias ayant un compte
Moins de 1 000	156	25,4%
1 000 à 10 000	203	33,1%
100 000 à 1 000 000	167	27,2%
Plus de 1 000 000	88	14,3%

Tableau 11. Nombre d'abonnés des médias sur Facebook

Nombre d'abonnés	Nombre de médias	Proportion des médias ayant un compte
Moins de 5 000	190	24,2%
5 000 à 20 000	184	23,5%
20 000 à 100 000	180	23,0%
100 000 à 500 000	133	17,0%
Plus de 500 000	97	12,4%

Tableau 12. Nombre d'abonnés des médias sur YouTube

Nombre d'abonnés sur YouTube	Nombre de médias	Proportion des médias ayant un compte

Moins de 1 000	268	45,0%
1 000 à 10 000	137	23,0%
10 000 à 100 000	103	17,3%
Plus de 100 000	87	14,6%

Tableau 13. Nombre d'abonnés des médias sur X

Nombre d'abonnés sur X	Nombre de médias	Proportion des médias ayant un compte
Moins de 1 000	161	19,9%
1 000 à 10 000	282	34,9%
10 000 à 100 000	242	30,0%
Plus de 100 000	123	15,2%

Les inégalités de visibilité entre les médias et la concentration de la visibilité sur les RSN sont encore plus fortes si on les mesure au nombre de vues des messages plutôt qu'au nombre d'abonnés. Cette information n'est disponible que dans le cas de YouTube (voir le tableau 14). Près d'un cinquième des chaînes totalisent moins de 10 000 vues et près des deux tiers moins de 1 million de vues. En revanche, une cinquantaine de chaînes, soit près d'un dixième des chaînes de la base, totalisent chacune plus de 100 millions de vues.

Tableau 14. Nombre de vues des chaînes YouTube

Nombre de vues de l'ensemble des vidéos de la chaîne	Nombre de médias	Proportion des médias ayant un compte
Moins de 10 000	98	16,3%
10 000 à 1 million	252	42,0%
1 million à 100 millions	204	34,0%
Plus de 100 millions	46	7,7%

Conclusion

Cette base de données a vocation à servir d'appui à diverses recherches impliquant le recensement de médias diffusant leurs informations sur les réseaux sociaux. Elle offre la possibilité de produire des échantillons de médias et des corpus de publications de taille très variée, mais aussi de comparer l'offre et la visibilité de médias à l'aide de statistiques descriptives, qu'elles soient univariées comme dans le présent data paper, ou multivariées en lien avec des questionnements et hypothèses plus précis. Mais il faut rappeler que cette base de données ne peut en aucun cas se substituer au travail de sélection à conduire pour de futures recherches et à son ajustement à des objets et problématiques particulières. Les médias en ligne, tout comme leurs comptes sur les RSN, se caractérisent en effet par des rythmes de création et

de disparition relativement élevés, qui nécessitent l'actualisation régulière des recensements opérés. Plus encore, les principes de sélection des médias retenus étaient en accord avec les ressources et finalités du projet Feeding Bias, et n'avaient pas vocation à recenser l'ensemble des sources sur l'actualité politique française, ni même les plus visibles d'entre elles. En se focalisant sur des organisations basées en France, reconnues comme des médias par des institutions centrales du champ journalistique, cette base exclut par exemple les comptes de journalistes, ceux de médias étrangers très suivis en France, ou encore de très nombreux et divers producteurs d'informations sur l'actualité non reconnus comme journalistiques (par exemple les comptes de partis et de think-tanks, ou ceux d'intellectuels et idéologues de tout type). Enfin, le travail d'identification de médias, aussi long et répétitif soit-il, est une manière de se familiariser avec le champ journalistique autant que de gagner en réflexivité quant à ses propres (pré)notions de ce qu'est un média.

Bibliographie

Cagé, J., Hervé, N., & Viaud, M.-L. (2017) *L'information à tout prix*. Paris : INA.

Cointet, J.-P., Cardon, D., Mogoutov, A., Ooghe-Tabanou, B., Plique, G., Morales, P. (2021) Uncovering the structure of the French media ecosystem. https://doi.org/10.48550/arXiv.2107.12073

Lyubareva, I., Marty, E. (2022). Vingt-cinq ans d'information en ligne : une exploration des transformations structurelles des médias. *Les Enjeux de l'information et de la communication*, 23(1), p. 5–14.

Marty, E., Rebillard, F., Pouchot, S., Lafouge, T. (2012) Diversité et concentration de l'information sur le web. *Réseaux*, 176, p. 27–72.

Marty, E., Ouakrat, A., Pacouret, J. (2022) De Valeurs actuelles à VA+ : l'appropriation des formats et des logiques des réseaux socio-numériques par un média d'extrême droite. *Quaderni*, 107, p. 99–122.

Mattelart, T. (2020) Comprendre la stratégie de Facebook à l'égard des médias d'information. *Sur le journalisme, About journalism, Sobre jornalismo*, 9, p. 24–43.

Pacouret, J., Bastin, G., Marty, E. (2024) L'espace social des réseaux sociaux. Une approche relationnelle de l'usage des plateformes numériques en France. *Sociologie*, 15, p. 119–146.

Sebbah, B., Sire, G., Smyrnaios, N. (2020) Journalisme et plateformes : de la symbiose à la dépendance. *Sur le journalisme, About journalism, Sobre jornalismo* 9, p. 6–11.

Sedel, J., (2021) Construire l'indépendance en label de qualité: Le travail de singularisation des éditeurs de presse en ligne « indépendants ». *Politiques de communication*, 16, p. 13–51.

Smyrnaios, N., Rebillard, F. (2019) How infomediation platforms took over the news: A longitudinal perspective. *The Political Economy of Communication*, 7.